UNIVERSITAT DE BARCELONA

Facultat de Biblioteconomia i Documentació

Gestión de datos de investigación oceanográficos: propuesta de un modelo para Brasil

Fabiano Couto Corrêa da Silva

UNIVERSITAT DE BARCELONA

Facultat de Biblioteconomia i Documentació

Programa de doctorat
Informació i Documentació en la Societat del Conoixement

Gestión de datos de investigación oceanográficos: propuesta de un modelo para Brasil

Tesis doctoral presentada por
Fabiano Couto Corrêa da Silva
para optar al título de doctor por la Universidad de Barcelona

Director:
Ernest Abadal Falgueras

Barcelona, noviembre de 2016

Desde hace décadas se viene diciendo
que sabemos más de la luna que del fondo marino y aún es cierto

Michael Schulz

Sumario

RESUMEN

La gestión de datos oceanográficos es un tema complejo debido a la diversidad de formatos existentes, sistemas propios de estandarización de las bases de datos y políticas internacionales. Esta investigación propone el diseño de un modelo de gestión de datos aplicable a la situación brasileña. En el caso que nos ocupa, el modelo está orientado a permitir la organización de los datos científicos procedentes de las investigaciones marinas de Brasil y posibilita la interconexión entre centros de investigación del país y el intercambio con repositorios internacionales. El modelo propuesto plantea unificar los datos científicos en el campo de la Oceanografía brasileña tomando como referencia el modelo Arc Marine Common Marine Data Types, desarrollado por Wright y colaboradores (2007), permitiendo adaptarlo a las necesidades identificadas. Este diseño está fundamentado en una adaptación de los modelos internacionales teniendo en cuenta las necesidades apuntadas por los oceanógrafos brasileños que contribuyeron con nuestra investigación. Mediante el análisis del caso brasileño, a través de estudios de usuarios, fue posible identificar las principales carencias para la gestión de los datos oceanográficos en Brasil. El resultado de la investigación demuestra que existen diversos tipos de datos, de formatos de intercambio y de redes de cooperación con políticas de indexación propias. Com esta investigación presentamos un modelo conceptual que permite su aplicación directa en problemas relacionados con la gestión de datos oceanográficos en instituciones de investigación con escasez de medios, así como la adaptación del mismo por repositorios de datos en desarrollo.

ABSTRACT

Oceanographic data management is a complex issue because of the diversity of formats, proprietary systems standardization of databases and international policies. This research proposes the design of a management model applicable to the Brazilian situation data. In the present case, the is geared to allow the organization of scientific data from marine research in Brazil and enables interconnection between research centers in the country and exchange with international repositories. The proposed model poses unify the scientific data in the field of Brazilian Oceanography reference to the Arc Marine Marine Common Data Types model developed by Wright et al (2007), allowing adapt to the needs identified. This design is based on an adaptation of international models taking into account the needs targeted by Brazilian oceanographers who contributed to our research. By analyzing the Brazilian case through user studies, it was possible to identify the main shortcomings for the management of oceanographic data in Brazil. The result of the research shows that there are various types of data exchange formats and networks of cooperation with indexation own policies. Com this research we present a conceptual model that allows direct application in problems related to ocean data management in research institutions with limited means, and adapting it by developing data repositories.

ÍNDICE DE FIGURAS

INDICE DE TABLAS

ÍNDICE DE GRÁFICOS

GLOSARIO

ADDS - Antarctic Data Directory System (SCAR)

AMD - Antarctic Master Directory

BODC - British Oceanographic Data Centre

CCAD - Commitee for the Coordination of Antarctic Data (SCAR)

CNDP - Centro Nacional de Datos Polares

COI - Intergovernmental Oceanographic Commission (UNESCO)

COMNAP - Council of Managers of National Antarctic Programs

CPE - Comité Polar Español

DIF - Directory Interchange Format (SCAR)

EMODNET - European Marine Observation and Data Network

ESEOO - Establecimiento de un Sistema Español de Oceanografía Operacional

DMAC IOOS - Data Management and Communications (EEUU)

DNA - Designated National Agency

GEO - Group on Earth Observations

GEOSS - Global Earth Observation System of Systems

GMES - Global Monitoring for Envieonment and Security (UE)

GOOS - Global Ocean Observing System

IBIROOS - Iberia Biscay Ireland Operational Oceanography System

ICES - International Council for the Exploration of the Sea

ICSU - International Council for Science

IDE - Infraestructura de datos espaciales

IDECSIC Infraestructura de datos espaciales del CSIC

IDEE - Infraestructura de datos espaciales de España

IDN - International Directory Network (SCAR)

IFREMER - Institut français de recherche pour l'exploitation de la mer

IGME - Instituto Geológico y Minero de España

IGN - Instituto Geográfico Nacional (España)

IODE - International Oceanographic Data and Information Exchange

ISI - Ingénierie des Systèmes InformatiqueS (Francia)

IOOS - Integrated Ocean Observing System (EEUU)

JCADM - Joint Committee on Antarctic Data Management (SCAR)

JCOMM - Joint WMO/COI Tehcnical Commission for oceanography and Marine **Meteorology -** JRC Joint Research Laborarory (UE)

NADC - National Antarctic Data Centre

NOAA - National oceanographic and Atmospheric Administration (EEUU)

NODC - National Oceanographic Data Centre

OGC - Open Geospacial Consortium

SAIDIN - Satellite Image database Interface (CSIC)

SCOR - Scientific Committee on Oceanic Research

SCAR - Scientific Committee on Antarctic Research

SDC - Centros de Datos de Satélite

SeaDataNet - A Pan-European Infrastructure for Ocean and Marine Data Management

SGBD - Sistema de gestión de base de datos

SISMER - Systèmes d'Informations Scientifiques pour la Mer (Francia)

UNEP - United nations Environmental Programme

WMO - World Meteorological Organisation

WDC - World Data Centre

AGRADECIMIENTOS

Quiero expresar mi agradecimiento al Dr. Ernest Abadal Falgueras por la entusiasta labor de dirección de esta tesis.

La realización de este programa de doctorado no hubiera sido posible sin la concesión de una beca por el CNPq (Conselho Nacional de Desenvolvimento Científico e Tecnológico) de Brasil por medio del Programa Ciencias sin Fronteras.

De igual forma, este trabajo tampoco hubiera sido posible si no hubiera contado siempre con el apoyo y la confianza de la *Universidade Federal do Rio Grande* (FURG), que concedió el permiso para que pudiera efectuar mis estudios de doctorado.

A Dra. Remedios Meleros, por brindar su granito de arena.

La ayuda del Institut de Ciències del Mar de Barcelona, en nombre de Jordi Sorribas, Joan Olive, Oscar Chic y Emilio Garcia, durante el análisis de la situación de los datos de investigación que realizamos en el campo de la oceanografía, nos ha permitido una mayor comprensión de los resultados de nuestro estudio.

Y para finalizar, mi gratitud a todas las personas que ofrecieron apoyo y amistad, con quien conviví y aprendí durante estos años, muchas gracias.

1 INTRODUCCIÓN

1.1 Los datos de investigación

Actualmente presenciamos una gran transformación en el desarrollo de la ciencia y tecnología, debida a la impactante cantidad de datos que se producen en el desarrollo de las actividades de investigación científica, la mayoría en formatos digitales. Además, el progreso alcanzado a día de hoy en la transmisión de información ha alcanzado una velocidad que es compatible con el volumen de información que se produce y se consume. Si hasta ahora los investigadores en el transcurso de su trabajo necesitaban localizar documentos en distintas fuentes como repositorios y bases de datos, hoy requieren también los datos de las investigaciones (CONICYT, IDER, 2010). Por consiguiente, a la gestión del almacenamiento de los resultados científicos, hay que añadir ahora la creciente demanda por parte de los investigadores del acceso a los datos mismos.

Esto se ha debido fundamentalmente a que la tecnología digital se ha convertido en un elemento cada vez más omnipresente en los procesos de construcción del conocimiento científico, ya sea mediante el aumento de la capacidad de los instrumentos científicos, mediante la reconstrucción de la realidad a través de la simulación, o mediante la apertura de nuevas formas de colaboración para compartir datos de investigación. Asimismo, los avances en las herramientas de preservación y procesamiento de datos han abierto nuevas aplicaciones para las fuentes básicas de una investigación -los datos- dando así un nuevo impulso a las investigaciones científicas. Un informe de la Organización para la Cooperación y el Desarrollo Económico (OCDE) de 2007 ya hacía hincapié en este hecho, destacando algunas situaciones en las que los datos de investigación se convierten en un factor esencial: la cadena de la innovación, la cooperación internacional, la promoción de nuevas investigaciones, la formación de nuevos investigadores y, sobre todo, la promoción de actividades científicas más abiertas y transparentes, que tengan como principio la producción de conocimiento abierto, a disposición del público.

El vicepresidente de Oracle Europa Malhar Kamdem, asegura que los datos tendrán en el siglo XXI "el mismo poder que tuvo la electricidad en el siglo XX"[1].

[1] Conferencia realizada en Barcelona durante el evento BigDataCoe. Disponible en: < Oracle y Barcelona Digital abren un centro 'big data' de referencia europea >. Acceso en: 18 sep. 2015.

La excomisaria europea de la Agenda Digital Neelie Kroes, ya advertía hace años que los datos brutos son "el nuevo petróleo". Pero para refinarlos, advierte Kamdem, se necesitarán profesionales especializados. Estructurar los datos de manera adecuada es una necesidad para el progreso científico, en consecuencia con la diversidad de tecnologías y de métodos existentes de gestión de datos. Pero lo sorprendente no es sólo la cantidad de datos, sino también todo lo que podemos hacer con ellos, pues los nuevos avances en minería y visualización nos proporcionan nuevas formas de extraer información útil a partir de conjuntos cada vez más grandes de datos.

Como ha dicho Tim Berners-Lee[2], director del Consorcio de la World Wide Web, "los datos son preciosos y van a durar más que los propios sistemas". Tiene razón, aunque cabría añadir que los datos serán tan útiles como su plan de almacenamiento. Los que se encuentran dispersos en PCs, *notebooks*, *tablets*, móviles u otros dispositivos personales suelen ser difíciles de manejar y susceptibles a pérdidas. El acceso online permite por el contrario presentar los resultados de una investigación de manera más amplia y completa, lo que representa un enorme potencial para el avance científico. En concreto, facilita tanto la recopilación de los resultados de las investigaciones científicas como la aplicación de datos antiguos en contextos nuevos. Siendo así, no es de extrañar que el intercambio de datos de investigación esté encontrando un amplio apoyo entre los actores académicos. La Comisión Europea, por ejemplo, ha proclamado que el acceso a los datos de investigación aumentará la capacidad de innovación de Europa (European Commission, 2012). Al mismo tiempo, asociaciones nacionales de investigación están uniéndose para promover el intercambio de datos en el mundo académico. En este sector, el Knowledge Exchange Group (2015), un esfuerzo conjunto de las cinco principales agencias europeas de financiación, es un buen ejemplo de proyecto transnacional para fomentar la cultura del intercambio y la colaboración. Por otro lado, revistas como *Nature, PLoS One* o *Atmospheric Chemistry and Physics* adoptan cada vez más políticas de intercambio de datos con el objetivo de promover su acceso público.

---

[2] Entrevista Disponible en: < http://www.bcs.org/content/ConWebDoc/3337 >. Acceso en: 19 sept. 2015.

Podemos preguntarnos qué es lo que hace a los datos ser tan valiosos. Una respuesta en pocas palabras sería que si el conocimiento es el motor del avance científico, los datos son su combustible. Para los investigadores, una gestión adecuada de los datos de investigación permite nuevas formas de comparación y de descubrimientos, es decir, permite generar nuevos campos de investigación. Estamos pues ante el comienzo de un cambio de paradigma: empezamos a ser capaces de almacenar, procesar y analizar enormes cantidades de datos y esto puede cambiar la forma en la que llevamos a cabo las investigaciones o tomamos nuestras decisiones.

Nuevas investigaciones en el campo de la oceanografía generan datos que necesitan ser organizados y entendidos, por lo que requieren de un proceso continuo que identifique las dimensiones, categorías, tendencias, patrones y relaciones, revelando su significado. Este proceso es complejo e implica un trabajo de reducción, organización e interpretación de datos, que se inicia previamente en la fase exploratoria y que continúa durante todo el ciclo de investigación.

Una pregunta clave es de qué manera pueden los investigadores apoyar este proceso. Hay dos formas fundamentalmente. La primera es organizando los datos de manea adecuada en relación a su “ciclo de vida”. De hecho, una buena planificación desde el momento de la recogida de datos hasta su almacenamiento es la única forma segura de garantizar la continuidad de nuevas investigaciones. La segunda es el depósito en repositorios y revistas que proporcionen una infraestructura adecuada para indexar y recuperar conjuntos de datos.

Hay muchas áreas de actuación donde los datos de investigación oceanográfica son importantes, así como existen diversos grupos de personas y de organizaciones que pueden beneficiarse de esta disponibilidad, incluyendo las administraciones públicas. Asimismo, es imposible saber exactamente cómo y dónde serán mejor utilizados y valorados los datos abiertos, ya que surgirán con toda probabilidad nuevas formas de uso en el futuro.

Un informe de gran repercusión publicado en 2010 bajo el título *“A Surfboard for riding the wave: how Europe can gain from the rising tide of scientific*

data" ("Subir a la cresta de la ola: cómo se puede beneficiar Europa de la creciente marea de datos de investigación"), da a conocer las valoraciones de un grupo de expertos convocados por la Comisión Europea para evaluar los posibles beneficios de la puesta en marcha de una e-infraestructura global de datos. Esta sería muy importante porque permitiría a investigadores procedentes de diferentes áreas de conocimiento compartir datos y reutilizarlos. El informe parte de la constatación de que el avance de las nuevas infraestructuras para la gestión de los datos de investigación puede acelerar descubrimientos y cambiar la manera de realizar las investigaciones, creando un escenario completamente nuevo. La razón de esto es evidente: la revolución digital ha hecho que sea mucho más fácil almacenar, compartir y reutilizar datos. Los datos de investigación de todas las disciplinas están ahora casi universalmente en formato digital. El amplio acceso e intercambio de los mismos aumenta el retorno de las grandes inversiones que se realizan en investigación y tiene el potencial de hacer crecer de manera exponencial el conocimiento.

A diferencia de las formas tradicionales de archivo, con los datos de investigación no preocupa tanto el mantenimiento de registros por cuestiones legales, históricas o culturales, sino que sobre todo se intenta satisfacer las necesidades de los investigadores. La misión principal de un archivo de datos de investigación, por tanto, no es solamente la de conservar la memoria grabada de un grupo, organización o nación, sino sobre todo la de proporcionar un servicio de vital interés para la comunidad investigadora.

Pero para que sean útiles a la comunidad científica, los datos deben seguir una estructura y organización claras, y constituir colecciones informativas relacionadas y registradas en un formato adecuado al tema tratado, es decir, en el contexto de una deter- minada comunicación científica. De esta manera, de los resultados generados en una investigación se obtendrá un conjunto de datos que podrá almacenarse y ser reutilizado al distribuirse a otros investigadores, e incluso podrá ampliarse a áreas distantes a las de los objetivos iniciales de la investigación.

Desde esta perspectiva, analizaremos el proceso de registro de los datos de investigación y el papel de sus integrantes desde su recogida hasta su publicación. Para ello, las agencias de fomento para la planificación de los datos de investigación ofrecen una amplia gama de políticas de gestión a sus respectivas comunidades de investigación, incluyendo el acceso a catálogos y bases de datos a través de Internet, protocolos de retención, la creación de metadatos, la migración de datos a través de software y sistemas de hardware, y la formación y el desarrollo de las normas internacionales. Al ofrecer estos servicios, las agencias desempeñan un papel activo y estratégico en la formulación de nuevos métodos y técnicas dentro de los intercambios de datos y el desarrollo de nuevos estándares en todos los aspectos relacionados con su conservación.

Por otra parte los datos y la información están en constante movimiento, por lo que es difícil fijarlos en una forma estática permanente. La ciencia también está cambiando, ya que múltiples áreas del conocimiento negocian cómo se entienden los datos en todas las disciplinas y los dominios científicos. Los avances del conocimiento científico son aún más difíciles de establecer en una época en la que grandes conjuntos de datos están disponibles en todas las áreas. Es difícil replicar el intercambio de información entre colaboradores desconocidos, sobre todo entre distintas comunidades científicas y durante largos períodos de tiempo. Algunas transferencias pueden ser mediadas por la tecnología, pero muchas dependerán de la experiencia de los mediadores, ya sean investigador u otros actores involucrados en la gestión de los datos de investigación.

Las actividades de recopilación, creación, análisis, interpretación y gestión de los datos provocarán importantes cambios en la investigación científica. Thanos (2013) apunta que la “ciencia de datos” dará lugar a una nueva forma de organizar y llevar a cabo las actividades de investigación, hecho que podría desembocar en un replanteamiento de los enfoques a la hora de resolver problemas de investigación y de llevar a cabo descubrimientos fortuitos. La reciente disponibilidad de acceso a grandes cantidades de datos, junto con las herramientas avanzadas de análisis exploratorio como la minería y la visualización de datos, también comportarán cambios en la metodología

científica. Por ejemplo, uno de los puntos de vista que se viene planteando es que el método científico tradicional impulsado mediante hipótesis -básicamente un método deductivo-, se complementará cada vez más con un método basado en datos -esencialmente inductivo-.

Para explotar estos grandes volúmenes de datos, se necesitan nuevas técnicas y tecnologías. Para que estas sean de utilidad entre la comunidad científica, se requiere de una estructura y organización jerárquicas, deben constituirse colecciones relacionadas entre sí y registradas en un formato adecuado al objetivo por el cual se han recogido, así como deben ir acompañadas de un cuerpo descriptivo (los metadatos), que incluya, entre otras cosas, la autorización legal para acceder y difundir sus contenidos. Por ejemplo, los resultados obtenidos en las salidas de campo con el levantamiento de imágenes y cuestionarios ofrecen como resultado un conjunto de datos que se puede almacenar correctamente y ser reutilizado por otros investigadores de áreas diferentes a las de la investigación inicial.

Las posibilidades de reutilización de los datos se amplían cada vez más. En lo que se refiere al volumen, en la actualidad estamos escalando desde terabytes a zettabytes. Según Falcone (2011), en los próximos años la cantidad de información se multiplicará por 44, alcanzando los 35 zettabytes en 2020. Este autor estima que alrededor del 90% de los datos existentes en el mundo se crearán cada 2 años, o que el 80% de los datos en la actualidad no está estructurado y tan sólo el restante 20% está almacenado en bases de datos que permiten su análisis de forma estructurada.

Frente la necesidad de estructurar el creciente volumen de datos, en el proceso de gestión de datos de investigación, la confianza, fiabilidad y la facilidad de uso de los datos son fundamentales, pero en general aún es un reto implementar los requisitos de gestión necesarios. La adopción colectiva por la comunidad científica podría provenir de Horizon 2020, el nuevo programa del marco europeo de investigación e innovación para el período 2014-2020. Según la Comisión Europea, "un plan de gestión de datos es un documento que describe cómo los datos de investigación recopilados o generados serán tratados durante un proyecto de investigación y después de que se haya completado, describe

cuales serán recogidos siguiendo metodología y estándares específicos, y cómo serán compartidos, si serán abiertos y cómo se realizará la curaduría digital". El establecimiento de planes de gestión de datos aún es una realidad lejana para muchos países. Expresiones como "datos abiertos" o “*big data*" pueden ser populares, pero pocas instituciones tienen una política real de gestión de datos.

Es evidente que muy pocos investigadores se preocupan realmente por los registros de los datos de investigación: por lo general, los mantienen hasta que ya no los necesitan. Gracias a planes de gestión de datos, podrían asegurarse de que sus objetivos sean compatibles con la preservación y el acceso. Los datos de investigación son una parte esencial de las pruebas necesarias para evaluar los resultados de investigación y para reconstruir los hechos y procesos que conducen a dichos resultados. El valor de los datos aumenta a medida que se agregan en colecciones y están más disponibles para su reutilización. Pero solo llegaremos a entender este valor si nos movemos más allá de las políticas de investigación, de las prácticas y los sistemas de apoyo desarrollados en un momento determinado. Necesitamos nuevos enfoques para la gestión y el acceso a los datos de investigación puesto que al intentar desarrollar infraestructuras para ello, todas las partes deben trabajar en colaboración y ser sensibles a las necesidades de los investigadores y de los diferentes contextos en los que trabajan. También se deben tener en cuenta los avances técnicos y de formulación surgidos de políticas procedentes de los ámbitos nacional e internacional.

La gestión de los datos de investigación incluye procesos de preservación, uso y reutilización. El dominio de estos aspectos es fundamental para que los investigadores planifiquen su trabajo desde la concepción de su proyecto hasta la ejecución, uso y archivo. Nuestro objetivo es investigar el estado de la cuestión sobre las infraestructuras disponibles y las posibilidades de uso de los recursos en las diferentes áreas del conocimiento, para entonces proponer un modelo adecuado a las necesidades presentadas.

Para responder da demanda de Brasil, es necesario hacer la pregunta: ¿Es posible una infraestructura eficiente si no se comparten los datos de investigación oceanográfica?

Nuestra intervención reflexiona sobre este interrogante. Investigadores, gobiernos, agencias de financiación, empresas, universidades, etc. son los agentes involucrados en este proceso. Cada uno cuenta con intereses y roles que determinan cómo se están abriendo resultados de investigación a pares y a la sociedad en su conjunto.

1.2 Los datos oceanográficos

La comunidad científica oceanográfica es una de las más activas en la generación de datos procedentes de los estudios realizados por diferentes sectores relacionados con el medio marino, incluidos los académicos, militares, tecnológicos que poseen estructuras dispares, acordes a la diversidad de métodos de observación e instrumentación. Además, en la actualidad los datos oceanográficos son muy complejos y presentados en distintos formatos, como imágenes, textos, vídeos, sonidos, etc. Al mismo tiempo de poseer una relación compleja, los datos oceanográficos, agregados en diversos niveles (datos de sensores, datos históricos, etc.), tienen como característica un gran volumen de información.

Los objetos digitales de una colección oceanográfica presentan diversas características como la posibilidad de ser copiados indefinidamente sin que haya pérdida de su calidad por desgaste de manipulación y por tiempo, la utilización de poco espacio físico para almacenaje, pueden ser distribuidos por internet y recuperados remotamente. Por ello se considera importante incluir objetos digitales en la arquitectura de publicación y administración de datos, información y productos oceanográficos.

Los centros mantenedores de información oceanográfica trabajan conjuntamente con los administradores de datos marinos para tener disponibles datos que han sido analizados y reagrupados en forma de bases de datos electrónicas, bibliografías por Internet, depósitos regionales de investigaciones científicas almacenadas y accesibles, catálogos en línea de colecciones especializadas, o colecciones electrónicas de estudios científicos difíciles de encontrar.

Además del interés de la comunidad científica en los datos, su divulgación puede aumentar la reproducibilidad de los resultados de la investigación. Cuanto más investigadores hagan disponibles abiertamente sus datos en repositorios de acceso abierto, será mayor la probabilidad de que otros puedan replicar su trabajo, con beneficios evidentes para todos, puesto que la falta de reproducibilidad en los resultados de la investigación es un tema que preocupa no solamente a la comunidad científica, sino al sector privado, los gobiernos y a la sociedad.

Son diversos los beneficios que una infraestructura adecuada de gestión de datos oceanográficos pueden generar para este sector: por ejemplo, disponer de búsquedas más eficientes, una mejor asignación de los recursos para la investigación activa, mayores oportunidades para el intercambio y la reutilización de los datos.

El objeto de estudio, desde un punto de vista temático, se ha concretado en un estudios de usuarios junto a los investigadores brasileños y el análisis de las bases de datos y consorcios internacionales de datos marinos.

El examen de los repositorios de datos oceanográficos especificados se ha dirigido a descubrir la existencia de una estandarización internacional para el almacenamiento y intercambio del levantamiento de los datos primarios obtenidos en el entorno oceanográfico, incluso en la Antártida.

Han sido incluídos en el estudio los repositórios de datos oceanográficos que disponían informaciones básicas sobre la recolecta de datos, ya que la inexistencia de estos aspectos dificultan, y en ocasiones imposibilitan, la recuperación de la información en la red. El marco geográfico del estudio ha sido circunscrito a la revisión exhaustiva de repositorios de datos referenciadas por directorios que presentan levantamiento de los principales repositorios y consorcios en todos lo continentes. Dada el tamaño de dicha análisis, se presentaba inalcanzable la revisión de todos los repositorios de datos que lo conforman, siendo necesario siempre que posible el estudio global de los consorcios en nuestra investigación. Si cumplía con el objetivo de ofrecer un estado de la cuestión de los países considerados líderes en el desarrollo y

aplicación gestión de datos marinos, resultaba indispensable optar por unos estados bajo criterios especificados en la metodología de esta tesis.

En referencia a las entrevistas con los investigadores de Brasil, fueran escogidos los que presentan significativa contribución investigativa en publicaciones científicas oceanográficas y coordinación de laboratórios en universidades o órganos del gobierno brasileño.

1.3 Justificación

Estratégicamente el mar es vital para el desarrollo de la economía brasileña, porque es el medio por el que circula el 95% del comercio exterior de Brasil, que pasa por más de 40 puertos en las actividades de importación y exportación, además de las inversiones avanzadas para la extracción petróleo (MARINHA DO BRASIL, 2013).

Desde 1994, cuando Brasil se convirtió en signatario de la Convención sobre el Derecho del Mar en un evento patrocinado por las Naciones Unidas (ONU), entró en vigor un ajuste que obliga a los países interesados en la soberanía de la Zona Económica Exclusiva (ZEE) a realizar estudios sobre el medio marino. Asimismo no pueden prohibir el tráfico de los barcos pesqueros de otros países en la orilla del mar. Desde entonces, Brasil se ha comprometido a invertir en estudios marinos que tradicionalmente fueron relegados a núcleos aislados.

Las ciencias del mar en Brasil, no muestra un desarrollo acorde con su condición geográfica. Las características oceánicas del país, con zonas pesqueras, portuarias y de investigaciones oceanográficas, además de Territorio Antártico, le imponen la necesidad de un desarrollo mayor de una infraestructura de gestión e intercambio de datos resultantes de la investigación científica marina, tal que permita su conservación y accesibilidad integral por un número significativo de centros, institutos y universidades dedicadas a la investigación científico-marina, que se distribuyen por todo el país, así como la exportación para repositórios de datos internacionales.

En el caso brasileño, a menudo los datos marinos son el resultado de los proyectos con un taxonomía limitada, temporal y espacial de las investigaciones. Tomada aisladamente, los conjuntos de datos resultantes de estos proyectos son sólo de uso limitado en la interpretación de los fenómenos a gran escala. Más específicamente, no informan sobre una escala adecuada sobre las investigaciones oceanográficas.

Existen, no obstante, enormes carencias de desarrollo tecnológico para acometer con éxito la tarea de disponer acceso, a los datos científicos procedentes de la investigación nacional. Estas carencias se reflejan, entre otros aspectos, en la poca disponibilidad de bases de datos estructuradas y accesibles y, principalmente, en la la falta de infraestructuras para que los investigadores puedan disponer de los datos, ya que en su mayoría envían los paquetes de datos primarios a bases de datos internacionales. Aunque el país cuenta con el Banco Nacional de Dados Oceanográficos (BNDO), bajo supervisión de la Marina de Brasil, que es responsable de recoger los datos oceanográficos de las instituciones brasileñas e integrar la Comisión Oceanográfica Internacional (COI), un análisis de los mecanismos utilizados por parte de las unidades de investigación marina para divulgación de los datos científicos muestra una falta de estructura común e incompatibilidad entre ellos.

En cuanto a la infraestructura, está formada por los centros de investigación donde funcionan tanto las oficinas como los laboratorios debidamente implementados para el desarrollo de las diferentes áreas de las ciencias marinas. En muchos casos faltan recursos para ofrecer visibilidad, accesibilidad, permanencia de la información científica y técnica y aumento del impacto entre la comunidad científica por medio de los resultados de investigación. Han surgido como una respuesta de las instituciones, en especial las académicas, la necesidad de conservar, preservar y poner a disposición de su comunidad académica e investigadora, su patrimonio intelectual. Por lo tanto, la adopción de formatos de intercambio de datos y estandarización de los mecanismos de almacenamiento son necesarios para divulgar la ciencia oceanográfica brasileña internamente en el país y a nivel internacional.

A lo largo de la investigación, hemos asistido a foros donde se debatían la situación y las tendencias internacionales en materia de gestión de datos marinos. Ello permitió conocer las principales iniciativas y profusión de estándares que entonces circulaban –y aún circulan–, normalmente vinculadas a los diferentes ámbitos de interés (la Oceanografía física, biológica, química y geológica).

Nuestra investigación es de carácter aplicado y quiere ser una propuesta de política de desarrollo para la gestión de los datos oceanográficos en Brasil. Se espera que constituya un documento base de consulta que oriente e impulse adecuadamente los esfuerzos nacionales que permitan resolver los problemas más urgentes y prioritarios que requiere el almacenamiento y el flujo de los datos de la investigación marina.

1.4 Hipótesis de la investigación

A continuación se van a describir las hipótesis de partida que van a servir para articular nuestra a investigación, y que se pretenden constatar durante el desarrollo del presente trabajo

A pesar del esfuerzo gubernamental realizados por el Banco Nacional de Dados Oceanográficos (BNDO) desde el año 1994 para gestionar los datos oceanográficos en Brasil, no hay progreso en el desarrollo de una infraestructura en el ámbito nacional.

Actualmente, los centros de investigación en Brasil no trabajan de manera integrada, tampoco funcionan con una estructura común entre ellos, y por lo tanto no aportan ni demuestran cambios significativos en la estructura operacional.

Así pues, los procesos de gestión de datos oceanográficos que se han dado hasta el momento en los centros de investigación de Brasil son insuficientes, puesto que carecen de una estandarización efectiva, coordinada entre ellos y alineada con el escenario internacional. Esto dificulta que consoliden el papel fundamental que tiene la oceanografía de un país con elevadas extensiones costera y territorial.

1.5 Objetivos

A la vista de la situación expuesta, nuestra investigación se orienta a la búsqueda de soluciones para proporcionar un modelo de gestión de datos oceanográficos integrados que sea de utilidad para Brasil.

Se trata de definir un modelo para la difusión y reutilización de los datos procedentes de la investigación brasileña en estudios oceanográficos.

Para ello se han establecido cuatro objetivos específicos:

1) Situar marco teórico de los datos de investigación en general y de los datos oceanográficos en particular

2) Evaluar las iniciativas más importantes en la organización de los datos oceanográficos en el ámbito internacional.
Se va a llevar a cabo un estudio comparativo de los diferentes mecanismos de carga de datos así como de las estructuras de metadatos que se utilizan en las principales bases de datos oceanográficos existentes. De esta forma se dispondrá del marco de referencia fundamental para identificar la solución más adecuada a la situación en Brasil.

3) Analizar las principales lagunas del sistema de gestión de datos científicos en el ámbito de la oceanografía en Brasil.
Fue realizado una auditoría de información del funcionamiento de los sistemas de proceso y almacenamiento de datos científicos en Brasil para detectar sus principales limitaciones.
A la vez se han analizado los hábitos y las necesidades en el uso de datos de investigación por parte de la comunidad investigadora en oceanografía.
Se han identificado los hábitos y las demandas de la comunidad científica brasileña en lo que respecta al uso y acceso a los datos oceanográficos.

4) Proponer un modelo para la organización, gestión y reutilización de datos científicos procedentes de la investigación oceanográfica en Brasil.
Se trata de proponer recomendaciones para el desarrollo de un sistema de

gestión de datos en estudios oceanográficos generados por la comunidad científica brasileña, con el objetivo de resolver los problemas identificados en la evaluación del estado actual de la infraestructura.

1.6 Metodología

Se han utilizado diversos métodos y técnicas de investigación para llevar a cabo los objetivos antes indicados. A continuación indicamos los procedimientos adoptados para cada uno de los objetivos específicos de la tesis:

1.6.1 Evaluación de repositorios internacionales

En este apartado fue utilizada la metodología evaluativa, que se basa en el establecimiento de indicadores y su posterior aplicación a los objetos de estudio que se han seleccionado, en este caso, los repositorios de datos oceanográficos.

Los principales indicadores considerados fueron los diferentes formatos de registro, los sistemas de difusión de datos, las tecnologías utilizadas para la carga y la integración de datos oceanográficos, entre otros.

A partir de la consulta en directorios especializados y otras fuentes de información, se establecieron las bases de datos oceanográficos a nivel internacional y se procedió a su análisis para la aplicación del sistema de indicadores antes indicado.

En el capítulo 4 se encontrará información más detallada.

1.6.2 Auditoría de la situación en Brasil

Por medio de estudio cualitativo, a partir de una encuesta y entrevistas contactamos investigadores brasileños que trabajan con datos marinos a fin de conocer sus hábitos y necesidades en lo que respecta al funcionamiento del sistema actual de proceso de datos científicos en Brasil. Para eso, consideramos todas las etapas, desde la generación de estos datos a partir de los grupos de

investigación en oceanografía y estudios polares, hasta los procesos de almacenamiento, organización y difusión que se pudieran llevar a cabo.

Llevamos a cabo un diálogo con los principales centros brasileños oceanográficos de investigación y las universidades para recoger toda la información necesaria.

En relación con las normas utilizadas en Brasil, hicimos entrevistas y encuestas que demuestra una comparación exhaustiva entre los distintos sistemas de gestión de datos científicos que se presentan en esta investigación, incluido el análisis de las ventajas y desventajas, lo que permite establecer parámetros para la realidad brasileña. Para entender la problemática que esto causa, cuestionamos los entrevistados a respeto de las principales bases de datos oceanográficas en Brasil. Con el análisis presentado, fue posible identificar la forma cómo operan de forma integrada con universidades y gobierno. Desde el punto de vista tecnológico, interrogamos los entrevistados sobre como las bases de datos han adoptado formas de integrar centros de investigación creando conexiones entre todos los campos en desarrollo directo e indirecto.

En el capítulo 5 si puede encontrar información más detallada.

1.7 Estructura

La tesis se estructura en seis partes diferenciadas, que definen el proceso de análisis que se ha seguido en su elaboración:

- Marco teórico (capítulos 2 y 3): tiene por objetivo presentar el marco teórico de los datos de investigación. Pone énfasis en las directrices y políticas de retención e intercambio de datos para el avance de la discusión sobre formas para promover el uso de los datos de investigación. Además, presentamos una visión amplia de las infraestructuras internacionales utilizadas para la gestión de datos de los datos oceanográficos.

- Evaluación de la situación internacional (capítulo 4): presentamos un análisis de los principales repositorios de datos oceanográficos. Hemos establecido indicadores que tornaran posible hacer un análisis de los criterios utilizados por las principales plataformas para la gestión de los datos oceanográficos.

- Análisis de la situación en Brasil (capítulo 5): hemos dedicado un apartado, por medio de la aplicación de una encuesta y entrevistas, para tratar el análisis de los investigadores brasileños a respecto del escenario brasileño considerando el dominio que tienen sobre las alternativas de herramientas y normas para la preservación de los datos de investigación.

- Propuesta de modelo (capítulo 6): el modelo de datos Arc Marine abrió la posibilidad de concebir una propuesta de base de datos de modo integrado, y así surgió el producto de esta tesis: un modelo de gestión de datos oceanográficos para el escenario brasileño. Proponemos una base de datos de datos de investigación para almacenar los datos científicos recolectados por los investigadores, maximizando la oportunidad de compartir abiertamente esos datos y metadatos, no teniendo que volver a recolectarlos, ya que se pierde tiempo y es costoso. El modelo propuesto plantea unificar la producción científica brasileña en el campo de las ciencias del mar tomando el modelo Arc Marine Common Marine Data Types, desarrolladas por Wright y colaboradores (2007) por ser una propuesta de gestión de datos marinos avanzada y flexible que posibilita complementar las demandas adecuadamente al escenario que se plantea.

- Conclusiones: se intenta dar respuesta a las hipótesis y objetivos planteados, realizando un análisis de los repositorios internacionales y una valoración general del panorama brasileño después de aplicar la encuesta y el cuestionario.

2 DATOS DE INVESTIGACIÓN

2.1 El interés por los datos

A pesar de que los estudios, informes y declaraciones sobre gestión de los datos de investigación son todos ellos relativamente recientes, hay ya varias referencias destacadas. A continuación se comentan en orden cronológico.

En 2004 se reunieron en París los ministros de Ciencia y Tecnología de los países integrantes de la OCDE, conjuntamente con China, África del Sur, Israel y Rusia, para discutir sobre las directrices internacionales de acceso a los datos de investigación, y como resultado de la reuión se aprobó la *Declaration on Access to Research Data from Public Funding* (OCDE, 2004).

Ese mismo año se creó la *Open Knowledge Foundation* (OKF) para promover el acceso a contenidos y datos abiertos. Esta fundación ha puesto en marcha proyectos como *Comprehensive Knowledge Archive Network* o la iniciativa *Open Data Commons*, así como soluciones jurídicas para la apertura y reutilización de datos de investigación. En 2008 se creó la *Public Domain Dedication and License* (PDDL), una licencia pensada para el uso de bases de datos (Open Data Commons, 2008). Con la supervisión de la OKF, el Manual de Periodismo de Datos[3] nació en un taller de 48 horas encabezado por el *European Journalism Centre* y la *Open Knowledge Foundation* en la MozFest[4] celebrada en Londres en 2011. Posteriormente, la MozFest se amplió convirtiéndose en un proyecto de difusión internacional que actualmente cuenta con la participación de los principales representantes del periodismo de datos.

La OCDE quiso dar continuidad a la declaración de 2004 y mediante el *Committee for Scientific and Technological Policy* designó un equipo de expertos con el encargo de proponer un conjunto de normas para la promoción y el desarrollo de los datos de investigación procedentes de la financiación pública. Para ello

---

[3] Manual de Periodismo de Datos http://interactivos.lanacion.com.ar/manual-data.

[4] La MozFest es un evento anual organizado por la Fundación Mozilla que reúne interesados en discutir el futuro de la web y en compartir experiencias innovadoras.

contactaron con instituciones de investigación y representantes políticos de los países miembros de la OCDE y como resultado, en 2007 se presentó el documento *Principles and guidelines for access to research data from public funding* (OECD, 2007), unas directrices que facilitan el acceso a los datos de investigación generados con financiación pública. Los principios y las directrices contenidos en este documento pretenden orientar a los gobiernos, las organizaciones de financiación, las instituciones de investigación y a los propios investigadores. Sobre todo a estos últimos, les quieren ayudar a sortear los diversos obstáculos y desafíos surgidos a raíz del intercambio internacional de datos, como los problemas tecnológicos, de gestión institucional o financiera, asi como las cuestiones relacionadas con la financiación, producción, administración y uso de los datos.

En 2007 el *Research Information Network* (RIN, 2007) publicó un informe de las políticas y las prácticas vigentes en los principales proyectos de investigación del Reino Unido. Aunque las políticas en el momento del estudio se habían enfocado principalmente a la difusión de la investigación mediante artículos en revistas y actas de congresos, se vió que algunos proyectos de investigación tenían una buena infraestructura de curación de datos. Este estudio proporciona un panorama comparativo de lo que los diferentes grupos de financiación esperan de los investigadores que apoyan acerca de la gestión y acceso a los datos de sus investigaciones. Se examinan las políticas y la práctica de una selección de 25 de los mayores organismos financiadores de investigación, entre los cuales se incluyen los siete consejos de investigación del Reino Unido, siete universidades, una selección de departamentos del Gobierno, de organizaciones de investigación y de empresas que invierten significativamente en I+D.

También en 2007 la Comisión Europea publicó el informe *Scientific Information in the Digital Age* (European Comision, 2007), que examina la utilización de las tecnologías digitales para mejorar el acceso a las publicaciones de investigación así como la utilización de los datos como un importante motor de innovación. Propone poner en marcha un marco a nivel de la Unión Europea para apoyar nuevas formas de promover un mejor acceso a la información científica en línea y la preservación de los datos de investigación. En cuanto a medidas concretas, la

Comisión apoya la difusión en acceso abierto de los proyectos de investigación (mediante, por ejemplo, el reembolso de los costes de publicación) y ha destinado recursos para el desarrollo de infraestructuras de almacenamiento de datos y para la investigación en preservación digital.

Dos años más tarde, la Comisión Europea encargó a un grupo de expertos un informe sobre acceso, uso, reutlización y calidad de los datos de investigación, con vistas al año 2030. El objetivo principal era conocer los beneficios y los costes de la puesta en marcha de una infraestructura global de datos fiable y estable, que permitiera a los investigadores la utilización, reutilización y explotación de los datos de investigación de cara al máximo beneficio de la ciencia y la sociedad. El informe titulado *A Surfboard for Riding the Wave* no fue publicado hasta 2011, y está siendo utilizado la como referencia europea en la construcción de una infraestructura que maximice los beneficios del acceso a la información científica.

Aún en 2011, el Knowledge Exchange, una asociación con miembros de instituciones dedicadas a la creación de e-infraestructuras para la investigación y la enseñanza superior de cuatro países europeos, presentó una visión general de la situación actual respecto a los datos de investigación en Dinamarca, Alemania, Holanda y Reino Unido, proponiendo líneas generales para el desarrollo de una infraestructura de datos por medio de un programa de acción conjunta.

Otro informe de 2011 denominado *Riding the wave* (Van der Graaf; Waaijers, 2011), reuniendo la experiencia de expertos en el desarrollo de infraestructuras de datos, ofrece algunas recomendaciones dirigidas a los sectores público y privado para gestionar datos y generar al mismo tiempo beneficios para la sociedad y la ciencia. Señala que para que un plan tan ambicioso obtenga éxito se necesita la participación de todos los interesados de la comunidad científica. El informe también recomienda que el manejo de datos debería convertirse en una competencia académica básica. Se identifican cuatro factores clave para el éxito de esta infraestructura: 1) los incentivos, 2) la formación de investigadores, tanto en su papel de productores como de usuarios de las infraestructuras de información de datos, 3) la infraestructura técnica y de organización, y 4) la financiación de la infraestructura para los nuevos desarrollos en la logística de datos. El informe señala asimismo tres objetivos estratégicos a largo plazo: el intercambio de datos

será parte de la cultura académica; los datos logísticos serán un componente integral de la vida profesional académica; y la infraestructura fijará el ritmo, tanto a nivel operativo como financiero. El informe también señala que para la comunidad científica, todavía muy ligada a los sistemas tradicionales de difusión de la ciencia en libros y artículos, la sistematización de los datos está muy lejos de materializarse, aunque entre aquellos sectores menos tradicionales sea posible avanzar en lo que se denomina datos de investigación globales.

En este sentido, el informe *Science as an open enterprise* de la Royal Society (2012) aporta un importante estudio sobre el uso de la información científica, que afecta tanto a los investigadores como a la sociedad. No sólo identifica las oportunidades o desafíos de compartir y divulgar la información científica, sino también cuestiona la manera de conseguir exprimir el máximo potencial de los datos de investigación, todo ello con la finalidad de apoyar una investigación innovadora y productiva que reporte beneficios a la sociedad. Además, el informe reconoce las ventajas fundamentales del acceso abierto a los datos de investigación y toma en consideración las condiciones específicas en las que la "apertura" es más beneficiosa a la comunidad de investigación y a la sociedad en general.

En España se presentó en 2012 el informe *Depósito y Gestión de datos en Acceso Abierto* elaborado por el grupo de trabajo del repositorio Recolecta, que aporta consideraciones notables a tener en cuenta. Algunas de ellas hacen mención al diseño e implementación de la política de gestión de datos de investigación, otorgando especial atención a la situación del país con respecto a otros. Pero también son identificadas la variedad de tipos de datos y los actores implicados en su gestión (los repositorios institucionales y temáticos, las agencias de financiación, los centros de datos existentes, los investigadores y los expertos en la gestión de datos). Asimismo, se reflexiona sobre los aspectos económicos derivados de la creación de una infraestructura interoperable de gestión de datos.

El informe europeo *Towards better access to scientific information: Boosting the benefits of public investment in research* (European Commission, 2012) constata el hecho de que hasta ahora los resultados de la investigación científica se han difundido sobre todo mediante artículos y advierte que no hay una práctica bien

establecida para la publicación de los datos. De ello se concluye que los investigadores son a menudo reticentes a compartir sus datos con la falsa creencia de que otros pueden beneficiarse injustamente de su trabajo. Además, el proceso de preparación de datos para compartir es percibido como un esfuerzo intensivo y por otra parte la ausencia de incentivos para el intercambio de datos supone un importante obstáculo importante. El informe se ocupa de esta última cuestión en sus recomendaciones.

En 2013 se publicó el informe *European Landscape Study of Research Data Management* que integra los documentos generados por el Proyecto SIM4RDM, que es resultado de una investigación desarrollada por seis socios Europeos: JISC (Reino Unido), HEA (Irlanda), NIIF (Hungría), NordForsk (Noruega), CSC (Finlandia) y SURF (Holanda). El estudio presenta los resultados obtenidos en una encuesta sobre las actuaciones de los organismos de financiación, instituciones de investigación, organismos nacionales y editores de los estados miembros de la Unión Europea y de otros países para mejorar la capacidad y las habilidades de los investigadores en el uso efectivo de las infraestructuras de datos de investigación. También, incluye recomendaciones a las organizaciones para ayudar a sus investigadores con el análisis de los programas paneuropeos, las subvenciones internacionales existentes y las políticas de las instituciones. Y el informe también examina si tales programas o políticas incluyen las intervenciones de apoyo a los investigadores que incentiva la obtención de los conocimientos, las habilidades y el apoyo necesarios para la gestión de datos.

En este breve repaso a estudios e informes sobre la gestión de los datos de investigación vemos que las cuestiones técnicas son los elementos que aparecen con mayor frecuencia, aunque seguramente la actitud de los investigadores respecto a la divulgación de sus datos es una de las cuestiones más importantes, y a la vez, más polémicas (Borgman, 2012). Además de las preocupaciones sobre la idea de compartir libremente datos de sus investigaciones, muchos investigadores están muy poco dispuestos a dedicar el tiempo necesario a conservar correctamente sus propios datos de investigación, básicamente porque muchos no han recibido una mínima formación de cómo hacerlo. A pesar de que existen políticas de preservación y compartición de datos como las de la National Science

Foundation (2015), la American Geophysical Union (2013), y la US Office of Science and Technology Policy (2013), entre otras, no está claro aún si estas directrices motivarán suficientemente a los investigadores a cumplirlas en un futuro. Sin duda, este es uno de los retos principales en la gestión de los datos de investigación.

2.1.1 Políticas de retención e intercambio de datos científicos

Para las investigaciones financiadas con fondos públicos hay algunas tendencias generalizadas sobre los períodos de retención de datos.

En Reino Unido el tiempo de retención es bastante largo. No todos los consejos de investigación señalan un periodo determinado de retención de datos, pero los que lo hacen, a menudo indican un periodo de al menos diez años (Cambridge University, 2010). Por ejemplo, la política del Engineering and Physical Sciences Research Council (EPSRC) señala que los organismos de investigación se asegurarán de que los datos de investigaciones financiadas por la EPSRC sean conservados de forma segura durante un tiempo mínimo de 10 años a partir de la fecha que el investigador obtenga los datos o, si la obtención de los mismos fue realizada por varios investigadores, a partir de la última fecha en la que el acceso a los datos fue solicitado por un tercero (Engineering and Physical Sciences Research Council, 2014).

En Estados Unidos el tiempo mínimo de retención para la investigación financiada por el gobierno federal es de tres años, según lo establecido por la White House Office of Management and Budget (2013). Los registros financieros, documentos de apoyo, registros estadísticos y todos los demás registros de entidades no federales pertenecientes a un premio federal deben ser retenidos por un período de tres años a partir de la fecha de presentación del informe final de gastos o, para las concesiones federales que se renuevan trimestral o anualmente, a partir de la fecha de la presentación del informe financiero trimestral o anual, respectivamente (White House Office of Management and Budget, 2013).

En Australia el período de retención recomendado es de cinco años, según lo establecido por el National Health and Medical Research Council, el Australian Research Council y las Universidades de Australia en el Código Australiano responsable de la conducta de investigación. En general, el periodo mínimo recomendado es de cinco años a partir de la fecha de publicación. Sin embargo, en casos particulares el período debe ser determinado por el tipo específico de investigación. Por ejemplo:

- Para los proyectos de investigación a corto plazo que sirven solamente para fines de evaluación, como los proyectos de investigación realizados por estudiantes, es suficiente una retención de datos de 12 meses después de la finalización del proyecto.
- Para la mayoría de los ensayos clínicos, es necesaria la retención de datos de investigación durante 15 años o más.
- Para áreas como la terapia genética, los datos de investigación deben ser conservados de forma permanente (por ejemplo, registros de pacientes).
- Si se considera que el trabajo de investigación tiene valor para la comunidad o el patrimonio, los datos se deben conservar permanentemente, y con preferencia en una colección nacional (National Health and Medical Research Council et al 2007).

La tendencia general de tiempo de retención suele oscilar por tanto entre los tres, cinco y diez años a partir de la finalización del proyecto o publicación, en Estados Unidos, Australia y Reino Unido, respectivamente. Sin embargo es recomendable consultar las políticas del fondo de investigación en particular para conocer el punto de inicio y la duración exacta de retención obligatoria de los datos.

También es importante tener en cuenta que algunos tipos de datos pueden requerir períodos de retención más largos, como se ejemplifica en la referencia anterior del Código Australiano responsable de la conducta de investigación. En Estados Unidos la Office of Research Integrity señala que los datos de investigaciones confidenciales y patentes requieren períodos de retención más largos. También puede haber requisitos especiales en función del tema en cuestión. Por ejemplo, en el caso de investigaciones confidenciales con financiación de los National Institutes

of Health, se deben conservar los registros durante seis años después de la fecha final de la resolución del caso. Como se ha señalado antes, también es importante mantener los datos de investigación pertinentes a invenciones patentadas (Office of Research Integrity, 2014).

Al igual que en Estados Unidos y Australia, los períodos de retención más largos en Reino Unido son para determinados tipos de datos. El Medical Research Council (MRC), por ejemplo, exige períodos de retención más largos para los datos de la investigación clínica. Las expectativas del MRC para la retención de datos de la investigación son:

- Los datos de investigación deben conservarse durante un mínimo de diez años después de que el estudio se haya completado.

- Para la investigación clínica realizada en unidades e institutos de investigación del MRC, deben conservarse durante 20 años después de que el estudio se ha completado.

- Los estudios que proponen periodos de retención más allá de 20 años deben incluir una justificación válida, por ejemplo, los datos de investigación relacionados con los estudios longitudinales suelen ser retenidos indefinidamente y archivados y gestionados en consecuencia (Medical Research Council 2014).

Por lo tanto los datos médicos, de patentes o que son importantes para una investigación longitudinal suelen tener períodos de retención más largos que los datos normales.

Un último aspecto a tener en cuenta sobre las políticas nacionales es que a menudo las políticas de intercambio y de retención de datos están intrínsicamente unidas. En Reino Unido por ejemplo, hay un mayor énfasis en el intercambio de datos que en la retención de los mismos, lo que conlleva a períodos de retención más largos. La expectativa es que el intercambio de datos se produzca a través de un repositorio de terceros como soporte para una retención más larga. Para periodos de retención muy largos la recomendación es utilizar un repositorio que conserve sus datos al mismo tiempo que los pone a disposición del público. Otra opción es utilizar las estrategias contempladas en el capítulo 7 para gestionar los

propios datos, aunque esto no sea lo ideal para el intercambio de datos con largos tiempos de retención.

El lugar de trabajo es otra fuente común para las políticas de retención. Para los investigadores que trabajan en el sector industrial, la empresa es propietaria de los datos y por lo tanto va a determinar el período de retención. Algunas compañías tienen normativas explícitas y otras no, lo más importante es que la empresa sea responsable en última instancia del mantenimiento a largo plazo de los datos. Para los investigadores que trabajan en el ámbito académico en cambio, las situaciones son diversas. Algunas universidades tienen una política explícita sobre la retención de datos, pero la mayoría no. Un ejemplo es la política de Harvard:

"Los registros de investigación deben ser conservados, en general, por un período de no menos de siete años después del final de una actividad en los proyectos de investigación. Con este fin, un proyecto o actividad de investigación se considera como finalizado después de: (a) la presentación de informes finales al patrocinador de la investigación; (b) el cierre final de salida financiera de un premio de investigación patrocinado; (c) la publicación final de los resultados de investigación; (d) el cese de la actividad académica o científica en un proyecto de investigación específico, independientemente de si se publican sus resultados; lo que tenga lugar más tarde” (Harvard University, 2011).

Las políticas universitarias a menudo recogen las políticas de retención de datos nacionales, a veces requieren períodos de retención más largos y suelen especificar cuánto tiempo se deben conservar los datos. Sin embargo, estas políticas no son específicas para diferentes áreas del conocimiento, sino que señalan normas genéricas de preservación.

En relación a los requisitos de intercambio, los organismos de financiación son la principal fuente de consulta. En EE.UU. los National Institute of Health (NIH) y la Fundación Nacional de Ciencia (NSF) han sido los principales motores para el intercambio de datos, y los que comparten las políticas se han extendido a otras agencias federales de financiación. En 2013 la Oficina de Ciencia y Tecnología de la Casa Blanca publicó un memorandum sobre el acceso público (Holdren, 2013), requiriendo que los principales organismos de financiación federal promulgasen un

método para la gestión de datos incluyendo los requisitos de intercambio. Esta nota se ha tornado como parámetro a seguir en el intercambio de datos, además de ser un criterio para recibir fondos federales en Estados Unidos. En Reino Unido, Research Councils UK y Wellcome Trust han sido los principales impulsores de la política de intercambio de datos desde mediados de la década de 2000 (Wellcome Trust, 2010; Research Councils UK, 2011). Sus políticas son generalmente más fuertes que las de los financiadores de Estados Unidos, llamando a los "investigadores a maximizar la disponibilidad de los datos de investigación, con el menor número de restricciones posibles" (Wellcome Trust, 2010).

2.2 ¿Qué son los datos de investigación?

Intentar proporcionar una definición exacta de los datos de investigación es un reto, ya que implica adaptarse al contexto en el que se hace la pregunta. Los datos de investigación comprenden un área muy amplia (todas las disciplinas) y, por tanto, la definición puede variar de acuerdo con los planteamientos de los actores (los científicos, las instituciones, los financiadores, etc.) y con los propios contextos nacionales. En cualquier caso, una definición en términos generales puede ser:

Los datos de investigación son la información registrada o producida mediante cualquier forma o medios durante el transcurso de una investigación.

Los datos de investigación pueden ser de tipo numérico, descriptivo o visual y, reproducirse en formato papel (incluyendo notas de investigación en cuadernos, fotografías, etc.) o digital. El concepto de datos de investigación también hace referencia a las distintas herramientas como protocolos, códigos numéricos, gráficos y tablas que son necesarias para recoger y organizar los datos tanto en trabajos de campo como en el laboratorio. Incluyen no solo los materiales y muestras biológicas y/o ambientales extraídas sino también los resúmenes generados durante el transcurso de la realización de una investigación.

En resumen, según lo anterior, otra definición es:

Los datos de investigación son todas las evidencias que un investigador necesita para validar sus conclusiones tras una investigación.

En línea con lo que se ha señalado, varias entidades de difusión internacional como los *National Institutes of Health* (NIH) de EE.UU., la *National Science Board* (NSBD) y la *Organization for Economic Cooperation and Development* (OECD) coinciden en señalar los datos de investigación y que combinan con la definición de los datos de investigación como:

a) el material registrado durante el proceso investigador, reconocido por la comunidad científica y que sirve para certificar los resultados de la investigación que se realiza;
b) el material que proviene de una única fuente y es difícil, o bien imposible, obtenerlo de nuevo.
c) aquel que puede admitir muchas formas (textos, números, imágenes fijas o en movimiento, entre otras) con atributos o características que describan investigaciones y entidades.

Los datos de investigación son la materia prima generada a partir de la toma / recogida / observación de los fenómenos y sucesos de la realidad (números, caracteres, símbolos, imágenes, sonidos, ondas electromagnéticas, bits, etc). Constituyen los componentes básicos a partir de los cuales se crea la información y el conocimiento. Son generalmente de naturaleza representativa (por ejemplo, de mediciones de variables tales como la edad, la altura, el peso, el color, la presión arterial, la opinión, los hábitos, la ubicación de una persona, etc.). Pueden estar “implicados” o ser “derivados” (por ejemplo, los datos que se producen a partir de otros datos, como el cambio porcentual en el tiempo calculado mediante la comparación de los datos de dos períodos de tiempo). Por último, se pueden grabar y almacenar en formato analógico o digital.

Los datos de investigación no han sido considerados especialmente como un producto valioso hasta hace poco. Los investigadores se limitaban a difundir los documentos (libros, artículos de revista, etc.) propios de su actividad de investigación y mostraban las conclusiones sobre los datos que habían reunido y analizado. A lo largo del tiempo, se ha ido produciendo una evolución en la percepción del valor de los datos de investigación en todas las áreas del

conocimiento, en relación con una demanda que no ha parado de crecer. Así, durante los años cincuenta y sesenta del siglo XX los datos de investigación eran vistos solamente como un producto resultante de una investigación. En las siguientes décadas, no obstante, ya empezaron a ser considerados como un subproducto que podía servir como punto de partida para avanzar en nuevos estudios.

Desde finales de los años ochenta hasta el final de la década de los noventa el almacenamiento y procesamiento de información fueron los principales componentes de la cadena de gestión de datos de investigación. Fue entre 1990 y 2010 cuándo adquirió importancia la reutilización de los datos de investigación para facilitar el desarrollo de nuevas investigaciones. A partir de ese momento se origina un punto de inflexión y los datos pasan a ser tratados como sustrato, es decir, como un componente esencial para proporcionar nuevos avances científicos. Asimismo, se constató que algunos tipos de datos de investigación eran costosos de producir y por ende, los investigadores comenzaron a ver ventajas en la reutilización y el intercambio como métodos para alcanzar nuevas conclusiones o comprender mejor un área de estudio.

Actualmente, disponemos de tecnología adecuada para el procesamiento de datos masivos. Es lo que denominamos *big data*, un término que describe conjuntos de datos que son lo suficientemente grandes y complejos como para suponer un reto para las herramientas y las técnicas anteriores. La complejidad de los grandes datos se puede caracterizar por el volumen, la variedad y la velocidad en términos de frecuencia de medición. Hay muchas iniciativas disponibles en todos los sectores para mejorar la eficacia con la que los investigadores pueden llevar a cabo la recogida de datos, la integración y el uso preciso de la información procedentes de múltiples fuentes. Esta enorme cantidad de datos no se limita al almacenamiento en grandes servidores, sino que también es susceptible de ser procesada y analizada correctamente para obtener el mejor valor posible. En el campo científico, la captura, organización y manejo de estos grandes conjuntos de datos pueden favorecer la toma de decisiones basadas en la recolección de datos de forma rápida y con resultados efectivos. La relevancia de los datos en el contexto de las *big sciences* como astronomía, física y biología condujo no solo a la

aparición de nuevos modelos de investigación, sino también a la aparición de nuevos campos de estudio como la astroinformática o la biología computacional (Borgman, 2010).

Cada uno de los datos por sí mismo no proporciona información. Hemos pasado de la era de la generación de datos a la era del tratamiento de datos; por eso, para crear información sobre datos, necesitamos interpretarlos. Los datos pueden ser el comienzo (o la corroboración) de las ideas y con ellos a través de su uso y reutilización se puede seguir la cadena de producción del conocimiento:

Datos - Información - Conocimiento - Información - Datos

Figura 1: Tipos de contenidos y su evolución
Fuente: Cateriano (2010)

¿Para quién puede resultar útil estos datos? Está claro que los propios académicos los pueden reutilizar en sus investigaciones. Pero este interés va más allá. Las empresas se están aprovechando de estas posibilidades generando nuevos productos y servicios. Así, por ejemplo, las empresas que trabajan en el sector de la salud invierten mucho dinero para gestionar los datos, presupuestar el coste de obtenerlo, almacenar, procesar y recuperar datos, como registros médicos, contabilidad de costes y reclamaciones.

De hecho, la web está llena de aplicaciones basadas en datos. En el campo de la Oceanografía a menudo se utilizan boyas marinas como elemento constitutivo básico de una red de sensores inalámbrica para la monitorización del mar. Se describe el nodo sensor o mote que permite leer los datos de varios sensores oceanográficos y transmitirlos inalámbricamente hasta un servidor de datos accesible a través de Internet.

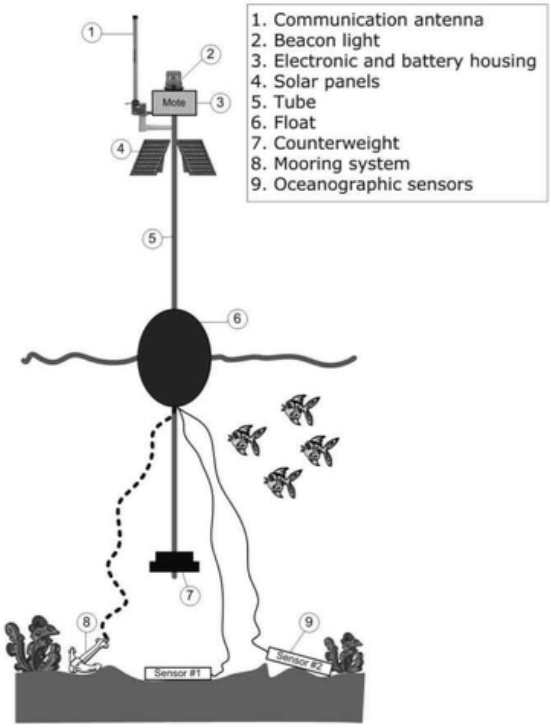

Figura 2: Estructura mecánica de una boya
Fuente: Albaladejo; et. al. (2011)

El mundo físico de la producción científica está determinado por los datos que dominan los procesos y productos. Las acciones y los productos en el proceso de fabricación son solo objetos físicos. Con datos precisos y oportunos, la ciencia puede avanzar con predicciones sobre la frontera del conocimiento en todas las áreas. En concreto, los datos y los mecanismos utilizados para su planificación son la fuerza impulsora que posibilita el crecimiento de la ciencia. Es inconcebible alcanzar el conocimiento sin datos, sin ellos no existiría el predominio de la estructura conceptual sobre los supuestos objetos.

En su esencia los datos son un mapeo del pasado que sirve para establecer relaciones con el conocimiento científico y alcanzar nuevos descubrimientos.

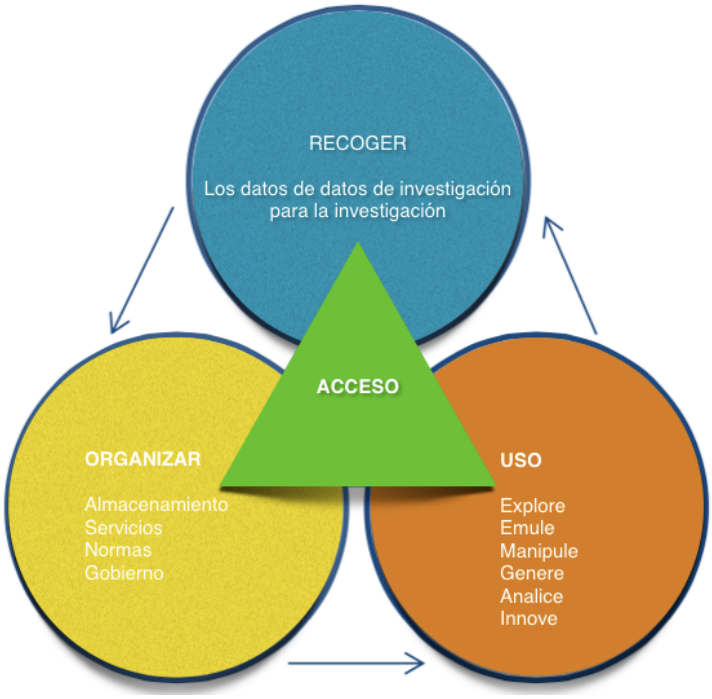

Figura 3: Elementos que componen un sistema de infraestructura de datos científicos
Fuente: Horizon 2020

Cuando una institución tiene un plan eficiente de gestión de datos, puede realizar sus actividades a un coste inferior, en menos tiempo, con menos recursos y mejores resultados que si no lo tuviera. Tomemos como ejemplo la medicina basada en evidencias (MBE), una metodología de localización, evaluación y utilización de datos de investigación utilizada en la medicina para las búsque- das realizadas por médicos para identificar los descubrimientos más recientes a nivel internacional. Por medio de bases de datos estructuradas se accede a conjuntos de informaciones que pueden agregar mucho valor en la atención médica, e incluso convertirse en la razón del éxito de tratamientos en enfermedades difíciles de diagnosticar. La estructura de la MBE se puede consultar en el programa VISC+,[1] una iniciativa de reutilización de datos de salud surgida para responder a las necesidades de la comunidad científica a través de información relacionada con la salud.

En casi todos los sectores se están utilizando nuevas tecnologías y herramientas de gestión de datos para llevar a cabo investigaciones científicas. Algunos proyectos combinan datos físicos y digitales, y los investigadores necesitan gestionar los datos de investigación para hacer un seguimiento de ambos a la vez. Con ello, los proyectos de investigación producen enormes conjuntos de datos que serían inmanejables sin la ayuda de ordenadores para procesarlos. Estas nuevas

tecnologías están abriendo las puertas a una mayor colaboración entre investigadores, informáticos y profesionales de la información en todos los campos de estudio, mientras que la gestión de datos no solo garantiza que estos puedan ser conservados y reutilizados posteriormente por terceros, sino también que los creadores puedan encontrar sus propios datos después de su uso inicial.

2.3 Tipología de los datos

Una primera distinción entre tipos de datos podría radicar en su carácter cuantitativo o cualitativo. Los primeros son cuantificables, es decir, se pueden expresar numéricamente. Los cualitativos, por su parte, son de carácter descriptivo e intentan reflejar la calidad o atributos de los objetos o sujetos, dando lugar a una categorización en vez de una cuantificación. Los investigadores de matemáticas, por ejemplo, a menudo son más propensos a usar datos cuantitativos, mientras que los investigadores de ciencias humanas son más propensos a usar datos cualitativos. La distinción entre estos dos tipos de datos no es tan evidente como puede parecer, aunque científicos de todas las áreas utilicen ambos tipos de datos en sus investigaciones. Más allá de esta distinción básica, según la National Science Foundation (NSF, 2007) los datos pueden ser clasificados mediante diferentes criterios: por el procedimiento de recogida, por su carácter primario o secundario, por el grado de estructuración, por su nivel de apertura y por su formato.

A continuación comentaremos cada una de estas categorías con la finalidad de obtener una panorámica más completa de la variada tipología de datos existente.

2.3.1 Según el procedimiento de recogida

2.3.1.1 Datos observacionales

Se trata de datos procedentes de observaciones científicas en las cuales los investigadores tratan de medir tantas variables como les sea posible a fin de dilucidar las posibles relaciones de causa y efecto. Son un tipo de registros que tienen un carácter único y que, a diferencia de los experimentos, no se pueden volver a reproducir y, por tanto, son insustituibles.

En definitiva son datos capturados en tiempo real, y por lo general fuera del laboratorio. Son resultado de registros de los hechos o las evidencias de fenómenos, generalmente con instrumentos. Los datos observacionales pueden ser por ejemplo observaciones climatológicas o de flora y fauna, por satélite, redes de sensores o mediante anotaciones por escrito. Pero también pueden ser en ciencias sociales, observaciones del comportamiento de distintos grupos o personas. Además, estas observaciones vienen asociadas a lugares y momentos específicos, o incluso a una multiplicidad de lugares y momentos (como en los estudios transversales y longitudinales).

2.3.1.2 Datos computacionales

Son los productos de la ejecución de modelos de ordenador, simulaciones, o flujos de trabajo. Los datos computacionales son reproducibles si se conserva la información sobre el modelo y su aplicación; es decir, para volver a utilizar un modelo computacional en el futuro es necesaria una amplia documentación del hardware y el software, de los datos de entrada y los pasos intermedios. Aunque son utilizados como parte del proceso de investigación en diversos campos, la bioinformática y la genómica han sido precursoras en el uso de datos computacionales con la detección de patrones y predicción de comportamientos. Su uso es más común en las ciencias físicas y de la vida, aunque también sean utilizados para el desarrollo de investigaciones en ciencias sociales y humanidades. Podemos citar como ejemplo los oceanógrafos, que utilizan datos computacionales para estudiar las mareas y hacer análisis sobre las probabilidades de aparición o predicción de tsunamis en determinadas regiones; o los economistas, que estudian las interacciones políticas y los mercados.

2.3.1.3 Datos experimentales

Son los resultados procedentes de experimentos, es decir, procedimientos realizados en condiciones controladas con el fin de probar o establecer hipótesis

sobre un determinado fenómeno. Si un experimento está diseñado para ser replicable, esos datos pueden ser más fáciles de reutilizar y preservar.

Por ejemplo, investigadores españoles han conseguido a través de experimentos obtener más datos que revelan las rutas de transporte de partículas entre el aire del interior y el exterior del ciclón conocido como vórtice polar ártico, aportando una ecuación que posibilita trabajar con aplicaciones matemáticas para el estudio de las corrientes oceánicas y atmosféricas[5] .

Varios tipos de registros están asociados con los datos observacionales, experimentales y computacionales, tales como los registros históricos, los registros de campo o las notas manuscritas. La grabación de audio es otro tipo que rara vez se define, a pesar de tener un extendido uso en el lenguaje diario. En este sentido, es como una cuarta categoría de origen de datos que abarca formas de datos que no encajan fácilmente en las categorías de observación, experimentación o computación. Los registros de casi cualquier fenómeno o actividad humana pueden ser tratados como datos posibles de investigación, incluyendo documentación del gobierno, los entes públicos y privados, documentación en forma de grabaciones de audio o vídeo, etc.

2.3.2 Los datos primarios, secundarios y terciarios

Según Cooper y Schindler (2003), los datos primarios son aquellos que han sido recogidos por un investigador a través de la realización de experimentos, encuestas, entrevistas u otras técnicas y que sirven para dar respuesta a un propósito específico, tratando de resolver el objetivo de la investigación. De esta manera, los datos primarios son trabajos originales de investigación y/o datos en bruto sin interpretación.

Si bien los datos secundarios pueden proporcionar mucha información, son menos precisos debido a que no fueron recogidos para las preguntas del inicio de la investigación. Los datos primarios se adaptan a los objetivos de investigación mientras que los secundarios no.

---

[5] Europa Press. *Matemáticos españoles revelan la ruta de partículas del vórtice polar antártico*. Madrid, jul. 2012. http://www.europapress.es/ciencia/noticia-matematicos-espanoles-revelan-ruta-particulas-vortice-polar-antartico-20120706133323.html

Los datos secundarios por su parte, están disponibles públicamente (en libros, publicaciones periódicas, censos, biografías, artículos y bases de datos, etc.) pero han sido recogidos con fines no estrictamente derivados del problema de investigación que nos ocupa. Esto significa que los datos primarios de un investigador se pueden convertir en datos secundarios cuando son analizados por una persona que no estuvo presente en el proceso de recogida y análisis de este material y que los usa para otra finalidad (Rabianski, 2003).

Los datos secundarios están disponibles para su consulta porque han sido recogidos, tabulados, ordenados, y puestos a disposición pública. La utilización de documentación en la investigación social constituye por lo tanto una fuente secundaria. Así, los datos primarios de una persona pueden ser datos secundarios de otra persona.

Los datos terciarios son una forma de datos derivados, tales como recuentos, categorías y resultados de datos estadísticos. A menudo los datos terciarios se utilizan para garantizar la confidencialidad de datos primarios o secundarios, cuando por ejemplo, en vez de mostrar nombres de personas entrevistadas, se presentan resúmenes de resultados. Quienes utilizan datos secundarios y terciarios como insumos para sus propios estudios tienen que confiar en que la investigación original es válida.

Los datos primarios siguen el método científico. Según el esquema clásico, primero se formula una hipótesis, a continuación se recogen datos y, tras su análisis, se discute y concluye si es correcta o no. Los datos secundarios en cambio, no empiezan a partir de una hipótesis, puesto que ya han sido recogidos.

El coste de la obtención de los datos primarios normalmente es mayor que el de los secundarios. Esto incluye coste económico y de tiempo. Los datos primarios incluyen los productos necesarios para desarrollar el experimento o recoger los datos, el análisis científico, y lo más importante, el tiempo para llevar a cabo la recopilación de datos, determinar los resultados, las conclusiones y luego redactar la información a publicar. Los secundarios suelen estar disponibles en repositorios, bases de datos estadísticas y publicaciones diversas.

Los datos primarios se utilizan en todos los campos de la ciencia, especialmente en las ciencias experimentales: la química, la biología, la física o la agricultura, entre otras. La gran mayoría de los artículos publicados en estos campos consisten en datos primarios. En ciencias sociales en cambio se utilizan más los datos secundarios. La investigación en esta área utiliza con frecuencia información secundaria para encontrar patrones en una zona de estudio.

En muchos casos los investigadores combinan datos primarios, secundarios y terciarios para producir datos derivados aun más valiosos. Por ejemplo, un investigador podría tratar de crear un conjunto de datos derivados que fusione sus datos primarios de análisis sobre los océanos con datos geográficos terciarios (sobre el tipo de peces que viven en diferentes áreas, que procedan de un estudio anterior y otros datos públicos y comerciales), con el fin de establecer recomendaciones para la preservación o la pesca. Los datos secundarios y terciarios son valiosos porque permiten la repetición de estudios y la creación de grandes conjuntos de datos, más ricos y más sofisticados. Más adelante se produce la "amplificación de los datos", es decir, los datos combinados permiten mayores conocimientos al revelar asociaciones, relaciones y patrones que estaban ocultos miemtras los datos permanecen aislados.

2.3.3 Según el grado de estructuración

Es importante diferenciar los datos que poseen una estructura de los que carecen de ella o la poseen parcialmente. Los datos estructurados se llaman así por estar almacenados de una manera estructuralmente identificable, mientras que los datos semiestructurados tienen el problema de falta de organización pues precisan de alguna manera que permita realizar consultas sobre ellos.

2.3.3.1 Datos estructurados y semiestructurados

Los datos estructurados son aquellos que pueden ser fácilmente transferidos a otros sistemas porque están organizados siguiendo un modelo de datos definido. Sería el caso de los caracteres alfabéticos o numéricos que figuran en las filas y

columnas de una tabla o base de datos relacional (por ejemplo, nombre, fecha de nacimiento, dirección, sexo, etc).

En los datos estructurados, la información está almacenada en tablas y las bases de datos están organizadas por un esquema, o sea, de una representación con su estructura, definiendo las tablas, los campos en las tablas y las relaciones entre ambos. Un ejemplo sería el propuesto a continuación:

Tipos datos no estructurados	Ejemplos
Imágenes	Satélite
Datos científicos	Gráficos sísmicos y atmosféricos
Fotografía y video	Grabación en los océanos
Acústico	Sónar y radar

Tabla 4: Datos estructurados
Fuente: elaboración propia

Por el contrario, los datos semiestructurados tradicionalmente incluyen imágenes, documentos de texto y otros objetos que no forman parte de una base de datos. No tienen un modelo de datos, un esquema predefinido y, por lo tanto, no se puede mantener en una estructura de base relacional. Por ser irregulares, son flexibles; a menudo añadidos jerárquicamente, pero tienen un conjunto razonablemente consistente de campos separados en contenidos semánticos y algún medio para ser clasificados y ordenados. Se trata por ejemplo de textos que incorporan un bloque de información que no encajaría tal cual en una base de datos. Ejemplos de este tipo de datos son las páginas web, que siguen ciertas pautas comunes y albergan contenido en el html y metadatos, entre las etiquetas.

2.3.3.2 Datos no estructurados

Los datos no estructurados no tienen un modelo de datos definido o estructura identificable corriente. Cada elemento individual, como texto narrativo o foto, puede contener una estructura o formato, pero no todos los datos dentro de un conjunto de datos de la misma estructura. Mientras que a menudo pueden ser buscados y consultados, no son fáciles de combinar o analizar informáticamente. Suelen ser

cualitativos en su naturaleza, y a menudo pueden convertirse en datos estructurados a través de la clasificación y categorización. Muchos análisis de *big data* se hacen sobre grandes conjuntos de datos semiestructurados o no estructurados, como los mensajes de Facebook, los tuits o los blogs.

Podemos entender mejor las diferencias entre datos estructurados y no estructurados si pensamos en la forma en que los investigadores describen determinadas áreas de investigación. Por ejemplo, las investigaciones en el campo de la Oceanografía son fundamentadas en impresiones subjetivas de los investigadores respecto a la naturaleza, a partir del análisis de agua de mar (datos no estructurados), con el objetivo de presentar datos estructurados que nos expliquen qué está sucediendo. Cuando un oceanógrafo define una observación del océano como “aguas moderadamente agitadas, recomendadas para pesca de peces del norte con el uso de embarcación mediana”, los datos no tienen una estructura subyacente (son datos no estructurados). La oración se puede cambiar, y no queda muy claro qué palabra se refiere a qué exactamente.

2.3.4 Datos abiertos

Los datos abiertos son los que pueden ser utilizados y distribuidos por cualquier persona, sin barreras técnicas o legales, bajo los requerimientos de reconocer la autoría y compartir el nuevo producto en las mismas condiciones.

Habitualmente, los organismos financiadores de investigación requieren que los datos producidos en el curso de un proyecto estén disponibles en acceso abierto. Para identificar los permisos de uso, existen varios tipos de licencias electrónicas estándares específicos para conjuntos de datos abiertos. Las licencias más utilizadas son las Open Data Commons y las Creative Commons.

Creative Commons tiene como propósito motivar a los creadores para que definan los términos en los que se pueden usar sus obras, con qué derechos y en qué condiciones. Por su parte Open Data Commons es una iniciativa promovida por la Open Knowledge Foundation (OKF) que comprende una serie de instrumentos jurídicos (licencias) para ayudar a generar y usar datos abiertos. En la tabla 2 se muestra un resumen de las licencias Creative Commons y Open Data Commons.

Tipo de licencia	Creative Commons License	Open Data Commons License	Descripción licencia
Dominio Publico	CCo	PDDL	Dedicado al dominio público. Renuncia a todos los derechos.
Atribución	CC BY	ODC-By	Permite cualquier explotación de la obra y de obras derivadas.
Atribución, compartir igual (share alike)	CC BY-SA	ODC-ODbL	Permite uso comercial, también de la obra derivada, bajo la misma licencia que la original.
Atribución, no comercial	CC BY-NC	-	Restringe el uso de actividades no comerciales.
Atribución, no derivados	CC BY-ND	-	Restringe la creación de trabajos derivados.
Atribución, no comercial, compartir igual (share alike)	CC BY-NC-SA	-	Debe proporcionar la atribución, reutilizar el contenido sólo con fines no comerciales, y colocar una licencia de compartir similar en trabajos derivados.
Atribución, no comercial, no derivados	CC BY-NC-ND	-	No se puede modificar el original o utilizarlo comercialmente y debe proporcionar la atribución.

Tabla 5: Licencias Creative Commons y Open Data Commons (2015)
Fuente: Adaptado de las recomendaciones de la Creative Commons y Open Data Commons

A continuación se describen pormenorizado las licencias disponibles a través de “Creative Commons (CC)” y “Open Data Commons (ODC)”:

CREATIVE COMMONS CERO (CC0)

Esta licencia disponibiliza información sobre los derechos de autor y dedica la información al dominio público. Está especialmente recomendada para datos científicos ya que facilita la compilación y el uso masivo de la información.

ODC “PUBLIC DOMAIN DEDICATION” (PDDL)

Esta licencia es equivalente a la CC0 y dedica la base de datos al dominio público. A diferencia de la licencia CC0 solo se aplica para bases de datos.

RECONOCIMIENTO (CC BY)

Asegura que siempre se de crédito adecuado a los autores, pero dificulta el uso simultáneo de varios conjuntos de datos al requerir que se siga una forma de citar predeterminada que puede ser diferente según el publicador.

ODC “ATTRIBUTION” (ODC-BY)

Es equivalente a la CC BY y asegura que siempre se dé crédito adecuado a los autores.

RECONOCIMIENTO COMPARTIR IGUAL (CC BY-SA)

Permite que independientemente de las modificaciones realizadas en la información original, esta siempre sea de uso libre. Sin embargo, no permite mezclar con facilidad la información pues puede haber incompatibilidad entre diferentes licencias “Compartir igual”.

ODC “OPEN DATABASE” (ODBL - DBCL)

Es una combinación de licencias, una para la base de datos (ODbl) y otra para los contenidos (DbCL), en conjunto son equivalentes a la CC BY-SA.

RECONOCIMIENTO SIN OBRA DERIVADA (CC-BY-ND)

Protege los intereses particulares del publicador, pero no permite generar nueva información a partir de la base de datos. No es acorde con la cultura de acceso libre.

RECONOCIMIENTO NO COMERCIAL (CC BY-NC)

Promueve el acceso a la información sin ningún coste, y es principalmente útil en el ámbito académico. La interpretación de la licencia puede ser subjetiva, a pesar de que actualmente no exista un acuerdo internacional acerca de lo que es un “uso comercial”. No es acorde con la cultura de acceso libre.

Al compartir los datos públicamente, el investigador debe optar por usar una licencia de derechos de reutilización que diferencie específicamente lo que está permitido y lo que no se puede hacer con los datos. Sin una licencia, quien consulta los datos tiene que averiguar si los derechos de autor se aplican al conjunto de datos y si tiene permiso de reutilización. Al existir diferentes restricciones a los derechos de autor en países distintos, la utilización de una licencia estándar es la forma más fácil de optimizar las condiciones de reutilización para cualquier persona en cualquier país.

Debido a que los derechos de autor no se aplican a todos los conjuntos de datos, la licencia recomendada para los datos de investigación es la de renuncia de dominio público (ya sea CCo o PDDL). Los datos relacionados con la ciencia publicada deben colocarse de forma explícita en el dominio público para maximizar su reutilización. Así, la renuncia de dominio público se está convirtiendo poco a poco en la norma para los datos compartidos, con muchos repositorios como Figshare, Dryad, BioMed Central (Cochrane, 2013), etc.

La única limitación que tiene un contenido de dominio público es que no puede haber un defensor legal de una licencia de atribución, ya sea CC BY o ODC-By, para los datos compartidos. Sin embargo, dada la ausencia común de derechos de autor, lo mejor es utilizar una renuncia de dominio público que se base en la obligación moral de requerir la citación al utilizar los datos de otra persona. Del mismo modo que hay que citar un documento cuando se ha consultado en una investigación, también se debe citar un conjunto de datos.

Otra ventaja de utilizar una renuncia de dominio público a través de una licencia de atribución es la combinación de múltiples conjuntos de datos (que a su vez han sido combinados de otros conjuntos de datos) con las múltiples licencias, para llevar a cabo un análisis más amplio. En cierto momento se hace difícil seguir los términos de licencia y los requisitos de atribución para cada único conjunto de datos o incluso determinar qué conjuntos de datos específicos de una base de datos son

parte del análisis. Por lo tanto la recomendación es que los investigadores utilicen una licencia de dominio público con la expectativa de que sean citados por la comunidad científica.

Aún cuando los datos están bajo derechos de autor,es recomendable una licencia abierta, como Creative Commons o Open Data Commons para compartir. La primera razón para elegir una licencia abierta es que el organismo financiador de una investigación o la política de acceso a los datos de una revista "esperan que todos sus investigadores financiados maximicen la disponibilidad de los datos de la investigación, con el menor número de restricciones posibles" (Wellcome Trust, 2010). Si bien esto no establece explícitamente el uso de una licencia, implica que sí debe utilizarse una licencia abierta para compartir los datos con restricciones mínimas. La segunda razón para usar una licencia abierta de datos con derechos de autor es que el intercambio de datos con restricciones de derechos de autor hace los datos disponibles, pero no reutilizables. El uso de una licencia abierta en todos los datos compartidos deja claro los permisos de reutilización en la presencia y ausencia de derechos de autor y maximiza el potencial de reutilización.

Por último, es importante destacar que incluso cuando los datos están legalmente en el dominio público, el autor de los datos puede determinar cuándo y cómo compartirlos. Una vez que haga disponibles públicamente estos datos, otros podrán reutilizarlos, pero el investigador siempre puede evitar su reutilización prematura por otros (por ejemplo, al ser consultado antes de la publicación de un artículo, en casos que no esté de acuerdo con las condiciones del uso de los datos para dar secuencia a determinadas investigaciones o propósitos).

2.3.5 Formatos de archivos de los datos

El formato y el software con que se crean los datos de investigación suelen depender de la forma en que los investigadores los recogen y analizan, y a menudo están influenciados por las normas y las costumbres propias de cada disciplina.

La información digital está diseñada para que sea interpretada por programas informáticos. Es por naturaleza dependiente de un software, lo que hace que los

datos digitales estén siempre potencialmente en peligro por la obsolescencia del hardware y software con los que han sido tratados.

Así, uno de los grandes desafíos del mantenimiento de los datos a largo plazo tiene que ver con el formato de los archivos digitales. Hay una gran cantidad de formatos de archivo disponibles para los diferentes tipos de datos de la investigación y a menudo es necesario un software especializado para abrir cada tipo de archivo. La opción más segura para garantizar el acceso de datos a largo plazo es la de convertirlos a un formato estándar. De esta manera, la mayoría de softwares serán capaces de interpretarlos. Formatos abiertos o estándares son por ejemplo Open Document Format ODF, ASCII, formatos delimitados por tabuladores, valores separados por comas, XML, etc.

Dado que los paquetes de software se van actualizando, a veces sucede que la nueva versión de un determinado software no abre los archivos de las versiones anteriores. Asimismo, es una realidad de la industria del software la inestabilidad de muchos de sus productos, por lo que también puede suceder que tengamos archivos en formatos para los que necesitemos un paquete de software que ya no existe. Este tipo de situaciones planteará continuamente desafíos de acceso a archivos antiguos.

Para contrarrestar estos problemas se debe planificar para el futuro la elección de los formatos de archivo en los que almacenar los datos. Obviamente, lo más adecuado es escoger un buen formato de archivo al inicio de un proyecto para la recogida de datos, aunque muchos investigadores no tienen esa opción. Aún así, vale la pena convertir archivos importantes al final de un proyecto mientras se preparan los datos para el largo plazo.

La clave para elegir un buen formato de archivo es que sea abierto, estandarizado, bien documentado y de amplio uso, y se deben evitar archivos en formatos desconocidos o propietarios siempre que sea posible. Si los archivos solo se pueden abrir mediante uno o dos softwares y además no son gratuitos, habría que considerar emplear un tipo de formato de archivo diferente.

En la tabla siguiente ejemplificamos más ampliamente este escenario:

Tipo de datos	Los principales formatos aceptados para el intercambio, la reutilización y la preservación	Otros formatos aceptables para la preservación de datos
D**atos tabulares cuantitativos con metadatos ampliados** Un conjunto de datos con etiquetas variables, etiquetas de códigos, y los valores que faltan definir, además de la matriz de datos.	Formato portátil SPSS (.por). Texto y comando archivo delimitado ('setup') (SPSS, Stata, SAS, etc.) que contiene la información de metadatos. Un texto estructurado o que contiene información de metadatos margen de archivo, por ejemplo, Archivo XML DDI	Formatos propietarios de paquetes estadísticos, ejemplo SPSS (.sav), Stata (.dta) MS Access (.mdb/.accdb).
Datos tabulares cuantitativos con metadatos mínimos Una matriz de datos con o sin descripción de columna o nombres de variables, pero no otros metadatos o etiquetado.	Valores separados por comas (CSV) (.csv). Archivo delimitado por tabuladores (TAB) incluyendo texto delimitado de carácter determinado conjunto con instrucciones de definición de datos SQL .	Texto delimitado de caracteres de datos - únicos personajes que no están presentes en los datos- se debe utilizar como delimitantes (.txt). Formatos ampliamente utilizados, por ejemplo, MS Excel (.xls / .xlsx), MS Access (.mdb / .accdb), dBase (.dbf) y Hojas de cálculo de OpenDocument (ODS). Texto con caracteres delimitados (.txt).
Datos Geospaciales Datos vectoriales y raster	ESRI Shapefile (esencial - .shp, .shx, .dbf, opcional - .prj, .sbx, .sbn). TIFF geo-referenciada (.tif, .tfw). Datos CAD (.dwg). Datos de atributos GIS tabular. Keyhole Markup Language (KML) (.kml).	Formato ESRI Geodatabase (.mdb). MapInfo Interchange (.mif) para los datos del vector. Adobe Illustrator (.ai), datos CAD (.dxf o .svg). Formatos binarios de paquetes GIS y CAD.
Los datos cualitativos Textual	eXtensible Markup Language (XML), utilizado de acuerdo con un tipo de documento apropiado (DTD) o esquema (.xml). Formato de texto enriquecido (.rtf). Los datos de texto sin formato, ASCII, (.txt), UTF-8 (Unicode).	HyperText Markup Language (HTML) (.html). Formatos ampliamente utilizados, por ejemplo, MS Word (.doc / .docx). Algunos formatos propietarios/específicos de software, por ejemplo, NUD * IST, NVivo y ATLAS.ti LaTeX (.txt).

Tipo de datos	Los principales formatos aceptados para el intercambio, la reutilización y la preservación	Otros formatos aceptables para la preservación de datos
Datos de imagen digital	TIFF versión 6 sin comprimir(.tif)	JPEG (.jpeg, .jpg) pero sólo si se crea en este formato. TIFF (otras versiones) (.tif, .tiff) Adobe Portable Document Format (PDF/ A, PDF) (.pdf). Norma RAW applicable al formato de imagen (.raw). Archivos de Photoshop (.psd).
Datos de audio digital	Free Lossless Audio Codec (FLAC) (.flac). Formato de forma de onda de audio (WAV) (.wav) Waveform Audio Format.	MPEG-1 Audio Layer 3 (.mp3) pero sólo si se crea en este formato. Audio Inter change File Format (AIFF) (.aif).
Datos de video digital	MPEG-4 (.mp4)	JPEG 2000 (.mj2)
Documentación y Plan de gestión de los datos	Formato de texto enriquecido (.rtf) Rich Text Format. HTML (.htm, .html). OpenDocument Text (.odt).	Texto (.txt). Algunos formatos propietarios ampliamente utilizados, por ejemplo, MS Word (.doc / .docx) o MS Excel (.xls / .xlsx). XML marcado de texto (.xml) de acuerdo con un DTD o esquema adecuado, por ejemplo, XHMTL 1.0 PDF (.pdf).

Tabla 6: Formatos de archivos
Fuente: Adaptado de las recomendaciones de la UK Data Archive (2015).

Es necesario el uso adecuado de formatos de archivos y de software para garantizar que los datos puedan ser identificados de acuerdo con las normas internacionales de forma única, y que sean accesibles para usos futuros. Al seleccionar herramientas para almacenar datos, es muy importante por tanto prestar atención a los formatos de los archivos. Para fines de preservación, y siempre que sea posible, el uso de formatos de datos debe estar disponible en estándar abierto y en un formato fácilmente reutilizable (por ejemplo .txt en oposición al .pdf). Para garantizar la elección adecuada del formato de datos, UK Data Archive (2015) recomienda tener en cuenta los siguientes consejos:

- disponer de un software necesario para ver los datos (por ejemplo, v.3 SPSS; Microsoft Excel 97-2003);

- tener información sobre el control de versiones;

- si los datos se almacenan en un formato en concreto durante la recogida y el análisis y luego fuesen trasladados a otro formato para su conservación, es recomendable hacer una lista de las características que se pueden perder en la conversión de datos, tales como etiquetas específicas del sistema.

Para una visualización resumida y esquemática, la tabla 3 contiene varios ejemplos de formatos de archivo seguidos de sus principales alternativas. La lista de tipos de archivos incluye formatos que se pueden abrir por varios programas de software, y que probablemente seguirán leyendo archivos en nuevas versiones. No hay reglas estrictas para elegir un formato de archivo, prevaleciendo el juicio personal sobre las opciones disponibles. No obstante es recomendable optar por formatos que tengan un uso extendido con el fin de tener una mayor probabilidad de apoyo en el caso de que sea necesario, principalmente porque se puede perder información durante el proceso de conversión. Asimismo, es importante tener en cuenta que la elección del formato de archivo de los datos no es una decisión final. De todos modos, si se observa un cambio sobre la preferencia generalizada ya sea por el uso de un determinado formato de archivo o por el uso de un software en particular, o simplemente se ha decidido cambiar los paquetes del software en la institución, será necesario actualizar los datos antiguos a los nuevos formatos de archivo. Por ejemplo, Lotus Notes y WordPerfect fueron en su día reemplazados por Microsoft Word como la forma estándar para archivos de texto, por lo que los archivos de texto antiguos realizados con aquellos programas deben actualizarse a los programas estándar posteriores.

2.4 Ciclo de vida de los datos

Durante una investigación científica se recogen y almacenan datos mediante un proceso sistemático dividido en etapas que, cuando se ponen en un determinado contexto, generan significados amplios sobre el objeto estudiado y los medios con los cuales fueron obtenidos. Por lo tanto el proceso científico es una combinación de dos etapas:

1 - El proceso de investigación, en el que se consumen, producen, procesan e interpretan datos.

2 - El proceso de preservación de datos, que ofrece sostenibilidad para el desarrollo de nuevos procesos de investigación.

Las actividades clave en este último proceso son la recopilación, la simulación y el análisis de datos, que generarán los que se introduzcan en el proceso de preservación. La salida directa de este proceso es la publicación científica, que a su vez conduce a los resultados indirectos de impacto social y económico. Aunque esto no sea explícito en el modelo del proceso, hay que señalar que el camino hacia el impacto social y económico no tiene que pasar necesariamente por la publicación científica formal: la reutilización de los datos intercambiados por la industria o los responsables políticos pueden por sí mismos producir un impacto socioeconómico sin el acompañamiento de publicaciones científicas.

El proceso de investigación durante la recopilación y el análisis plantea el problema de que los datos deben ser conservados para permitir su intercambio y reutilización. En muchas disciplinas se recogen los datos en bruto y luego son tratados. Mediante el proceso de análisis se generan conjuntos de datos derivados en cada etapa, antes de que se produzca el resultado final. Los datos resultantes son por lo general los que se publican o archivan en las situaciones en que la preservación es un requisito del proyecto. Sin embargo, cada disciplina trata la recogida de estos conjuntos de datos de manera diferente, mientras que los resultados de etapas anteriores son requeridos a menudo.

El descubrimiento de los datos y el acceso a ellos ponen de relieve el papel potencial de los servicios complementarios de preservación digital, como la indexación de datos en los motores de búsqueda, los cuales pueden ser integrados por medio de muchos archivos de datos. Por ejemplo, los repositorios de datos Figshare y Zenodo se encuentran disponibles para todas las disciplinas; otros son especializados, como Dryad en bCOliencias, Pangaea en ciencias de la tierra, etc. Los servicios de búsqueda también pueden ser vinculados a otros servicios complementarios; por ejemplo, con el número de citas de artículos publicados con referencia al artículo. Hay muchas posibilidades en cuanto a la integración de los servicios para apoyar el descubrimiento de datos, lo que podría ser proporcionado

por varios de los actores del proceso. DataCite y CrossRef trabajan conjuntamente para asegurar que los investigadores puedan navegar sin problemas entre los resultados de una investigación haciendo que artículos y datos sean identificables, referenciables y citables en el expediente académico.

DataCite es un consorcio mundial que asigna códigos identificadores a la investigación de datos y la mejora al permitir encontrar, compartir, usar y citar esos datos. Es una organización internacional que se relaciona con las partes interesadas, incluidos investigadores, académicos, centros de datos, bibliotecas, editores y proveedores de fondos, mediante la promoción, la orientación y los servicios.

CrossRef por su parte funciona como un *hub* digital para la comunidad académica, proporcionando una amplia gama de servicios para la identificación y localización de paquetes de datos y artículos específicos de forma masiva, obteniendo la dirección asignada a cada uno de ellos para acceder directamente a la página de la editorial donde se encuentra el texto completo.

2.4.1 Ventajas de compartir datos

Las ventajas de compartir datos son amplias y han llevado a muchas organizaciones, entre ellas el International Council for Science (ICSU) y la COI, a la adopción de una política de acceso abierto a los datos. Los océanografos son un buen un ejemplo con el World Ocean Atlas (WOA), publicado por el National Oceanic and Atmospheric Administration (NOAA). Gran parte de nuestra comprensión de los patrones globales se basa en estas bases de datos (Levitus, 1996; Conkright; Levitus, 1996).

Para cualquier iniciativa en que sea necesario confiar en la buena voluntad de compartir datos, se ha de tener en cuenta la sociología de la ciencia: los investigadores tendrán que ver claramente las ventajas de compartirlos y necesitarán incentivos para hacerlo. Tendrán que ser compensados por el tiempo que pasen haciendo que los datos estén disponibles para su reutilización, por la pérdida del acceso exclusivo y de la ventaja competitiva asociada a este. Un caso claro de este tipo de incentivos es cuando los datos se comparten entre varios proveedores con la intención de analizarlos agrupados y de publicar los resultados

de manera conjunta. Algunos ejemplos son el North Sea Benthos Project of the International Council for the Exploration of the Sea (ICES) y otras iniciativas de la Unión Europea como la Marine Biodiversity and Ecosystem Functioning (MARBEF). El incentivo es la oportunidad de analizar un conjunto de datos más grande que el que está disponible a partir de un único proveedor y el poder ser coautor de los *papers* resultantes.

Los datos se recogen a menudo con financiación pública, por lo que muchos creen que solamente por esta razón deberían ser accesibles al público. En ocasiones existe una obligación contractual para que los datos estén disponibles después de la publicación de los resultados. Los organismos de financiación hacen inversiones en investigación para profundizar en muchas áreas del conocimiento científico y la retención de los datos brutos obstaculiza el proceso por el cual los resultados de las actividades financiadas se pueden utilizar, contraviniendo claramente la intención original del apoyo de estos organismos (Dittert; Grobe, 2001). Demasiados datos se mantienen inactivos, algunos en discos duros, otros en formatos electrónicos de difícil acceso y otros solo disponibles en papel. Los oceanógrafos físicos han dado un buen ejemplo con la Global Data Archaeology and Rescue (GODAR), gracias a la cual fueron recuperados e integrados muchos conjuntos de datos en riesgo de perderse. El coste de recuperación de datos es normalmente solo una fracción del coste de la recogida y generación de esos datos. Más importante incluso que estos argumentos económicos es el elemento histórico de algunos datos de investigación, que pueden ser insustituibles, ya que una vez perdidos no se pueden recoger de nuevo.

La sinergia entre revistas científicas y repositorios de datos ofrece importantes posibilidades para descubrir y reutilizar resultados de investigación, como el World Atlas of Marine Ecosystem Data (MAREDAT). Ante el reto de la investigación abierta y la falta de coordinación entre instituciones, una de las dificultades reside en la caracterización de los problemas específicos para compartir datos de investigación, porque están potencialmente implicados en diferentes aplicaciones al servicio de diversas comunidades de práctica y en el contexto de las infraestructuras de investigación. Es fundamental disponer de un entorno de desarrollo rico y flexible donde verificar en la práctica la efectividad de los enfoques en acceso abierto que ya existen (como la atribución de registros DOI a los datos).

Además, los resultados obtenidos en estudios de campo componen un conjunto de datos que pueden ser archivados adecuadamente y reutilizados por otros investigadores, incluso de áreas distintas o paralelas a las de los objetivos iniciales de la investigación. La *data curation* o curaduría de datos, añadiendo metadatos a los registros e incluyendo la autorización legal para acceder a ellos y reproducirlos, es determinante para la continuidad del uso de los datos de investigación. Por lo tanto, considerar la publicación de los datos de una investigación como un documento de datos tiene un gran valor para los demás. Algunas revistas, como PLOS ONE y Ecology aceptan trabajos de datos como uno de los tipos de artículos que publican, mientras que otras como Scientific Data y Open Health Data publican exclusivamente trabajos de datos.

Inicialmente es necesario garantizar que los datos estén registrados, mantenidos y preservados de manera adecuada. Uno de los primeros requisitos es que los conjuntos de datos estén acompañados de informaciones que describan su origen (tiempo o espacio, métodos e instrumentos de recogida), ámbito, autoría, propiedad y condiciones de reutilización, o sea, de metadatos. Así, juntamente con la interoperabilidad tecnológica, la existencia de metadatos adecuados y normalizados es un requisito esencial para el acceso y la reutilización de datos de investigación.

La curaduría de datos no termina en la creación de metadatos, ya que comprende también un conjunto de acciones para garantizar su autenticidad, integridad y accesibilidad, incluyendo todas las actividades de preservación necesarias para garantizar la posibilidad de su uso posterior. Llevar a cabo una gestión eficiente de una cantidad de datos inmensa es muy complejo y necesita una arquitectura adecuada para mantener un flujo de datos adecuado.

Dado el creciente número de repositorios institucionales, se ha sugerido que estos podrían ser la respuesta, o al menos parte de ella, a la necesidad de curaduría de los datos producidos en la investigación. Actualmente varias instituciones de investigación de todo el mundo están empezando a darse cuenta del verdadero valor de la preservación digital en el contexto de los datos de investigación. De hecho, ya hay varios repositorios institucionales en la etapa de producción, pero su uso para albergar, conservar y proporcionar acceso a los conjuntos de datos de investigación es aún muy reducido. La curaduría de los datos generados por la

comunidad científica es por tanto un desafío estratégico clave para los gerentes y administradores de repositorios institucionales. La existencia de la voluntad, las competencias y los recursos necesarios para hacer frente a estos desafíos en las instituciones aún debe ser plenamente demostrada.

Como los repositorios institucionales pueden ser responsables de la custodia de los datos de investigación, tendrán que desarrollar estrategias específicas para cada área, ya que un enfoque genérico de curaduría de datos no es suficiente para responder a todas las necesidades y expectativas de los investigadores de las diferentes disciplinas. Es precisamente esta necesidad de combinar la dimensión institucional (muy amplia y multidisciplinaria en el caso de las universidades) con la dimensión temática o disciplinar (con sus requisitos específicos), la que constituye un desafío importante para el uso de los repositorios institucionales como un componente clave en la estructura de curaduría general de los datos de investigación.

La LERU (League of European Research Universities) ofrece algunos casos que ilustran estas cuestiones. Por ejemplo, la Universidad de Oxford estima en dos millones de libras la implementación de los siguientes servicios: DataFinder (catalogación de datasets), DataBank (almacenamiento y preservación), coordinación, servicio de almacenamiento y otras bases de datos. Por su parte, el University College de Londres hace un estudio de costes en base a las unidades de almacenamiento y calcula en un millón de libras el establecimiento del servicio. Ambas tienen un coste anual de mantenimiento de medio millón de libras (LERU, 2013). La LERU representa el reconocimiento de que las universidades trabajan ahora en la era de la ciencia basada en datos y presenta estudios teniendo en cuenta el vínculo entre la política de datos de investigación, la tecnología y el soporte técnico. Pone en evidencia que para promover el uso consciente y exitoso de datos de investigación, estos tres aspectos se deben ofrecer de forma simultánea a los investigadores y destaca que los proyectos que solo se centran en uno o dos de ellos están condenados al fracaso, así como los que no se alinean con las políticas que favorecen los servicios de apoyo.

El Digital Curation Centre (DCC) en Reino Unido y el Danish Archiving and Networked Services (DANS) de Dinamarca son dos organizaciones que proporcionan apoyo a la gestión de datos y a sus respectivas comunidades de

investigadores con el objetivo de fortalecer la base de conocimientos en esta área. El DCC proporciona asistencia en la redacción de planes de gestión de datos, capacitación, estudios de casos y documentos informativos. El DANS presentó un servicio similar con la elaboración de un Plan de gestión de datos que puede ser utilizado por investigadores como una lista de verificación en las primeras etapas de un proyecto de recopilación de datos.

El proyecto Opportunities for Data Exchange (ODE), financiado por el 7PM e integrado por los miembros de la Alliance for Permanent Access (APA), apoya la inversión para la reutilización, el intercambio y la preservación de los datos. La ODE analiza las opiniones de los expertos y las percepciones de los investigadores sobre *data sharing* y ofrece informaciones sobre los mejores proveedores de datos de las disciplinas que actúan, identificando los formatos de datos y metadatos más adecuados para ser descubiertos y reutilizados. Diversos países han abordado el problema inicialmente desde la sensibilización de los actores involucrados en la generación de conocimiento científico, como son los investigadores, los centros de investigación y los gobiernos y las organizaciones privadas que financian políticas de ciencia, tecnología e innovación o participan en ellas. Conjuntamente con todos estos actores, están configurando una política de gestión de datos e información científica con acciones para preservar los datos de investigación, como un registro histórico para la investigación actual y futura, mediante recomendaciones que apuntan soluciones viables a las barreras tecnológicas e institucionales. Es el caso del UK Data Archive, que tiene la mayor colección de datos de investigación digitales de Reino Unido en ciencias sociales y humanidades. Con millones de conjuntos de datos relativos a estudios sociales, tanto históricos como contemporáneos, es un recurso vital para investigadores, profesores y alumnos. En este ámbito, proyectos como el Data Asset Framework, el Edinburg DataShare o el DANS también son significativos ejemplos de preservación de los datos de investigación digitales.

2.4.2 Modelos del ciclo de vida

Hay dos maneras de considerar el ciclo de vida de los datos científicos: desde la perspectiva de un investigador y desde la de un bibliotecario. La mayoría de los

servicios de gestión de datos están relacionados con el apoyo a las necesidades de los investigadores en todo el proceso de investigación. Hay papeles para los investigadores en la mayoría de las etapas de este proceso y cada etapa se hace más fácil con una buena planificación y gestión.

La importancia de los modelos de ciclo de vida es que proporcionan una estructura para la consideración de las muchas operaciones que deberán llevarse a cabo en un registro de datos durante su uso. Muchas acciones de curaduría se tornan más fáciles cuando se preparan por adelantado, incluso antes de la creación del registro o en el mismo momento. Por ejemplo, los datos científicos de un repositorio son más útiles cuando ya han sido aclaradas las acciones de conservación que envuelven los derechos y las licencias de los datos. Así las metodologías y los flujos de trabajo utilizados por los investigadores tienen más probabilidades de ser grabados con detalle durante la investigación.

Los modelos del ciclo de vida de los datos cubren su vida útil después de haber sido creados, analizados y listos para ser sometidos a un repositorio. La curaduría de los datos significa su gestión una vez que se han seleccionado para la preservación y el almacenamiento a largo plazo. Los pasos secuenciales del ciclo de vida de los datos son, básicamente, la recogida, evaluación y selección, el uso directo en investigaciones, la realización de acciones de preservación, el almacenamiento, el acceso a los datos de uso, su reutilización y su transformación. Hay acciones puntuales que pueden interrumpir el ciclo, como revaluar y borrar conjuntos de datos.

Del mismo modo, los datos en sí pueden generarse de dos maneras: creándolos o transformándolos a partir de conjuntos ya existentes. Algunos elementos clave deben ser considerados en cada etapa del ciclo de vida, incluyendo la planificación y la preservación. Los modelos de gestión de datos están destinados a ser utilizados como guías para la planificación, lo que no significa necesariamente que sean un conjunto de reglas a seguir paso a paso, pero sí pueden ser útiles para la elaboración de informes entre los investigadores y administradores.

Muchos investigadores han tratado el enfoque del ciclo de vida para la gestión de activos digitales y han presentado modelos de ciclo de vida específicos. Higgins (2007), miembro del DCC (Digital Curation Centre), desarrolló el Curation Lifecycle

Model, que trata de manera específica las necesidades relacionadas con la “digital curation”. El modelo define un método de gestión documental basado en el ciclo de vida de los datos con el objetivo principal de alinear las tareas de curaduría con las etapas del ciclo.

Una representación gráfica del modelo puede ser vista en la Figura 4:

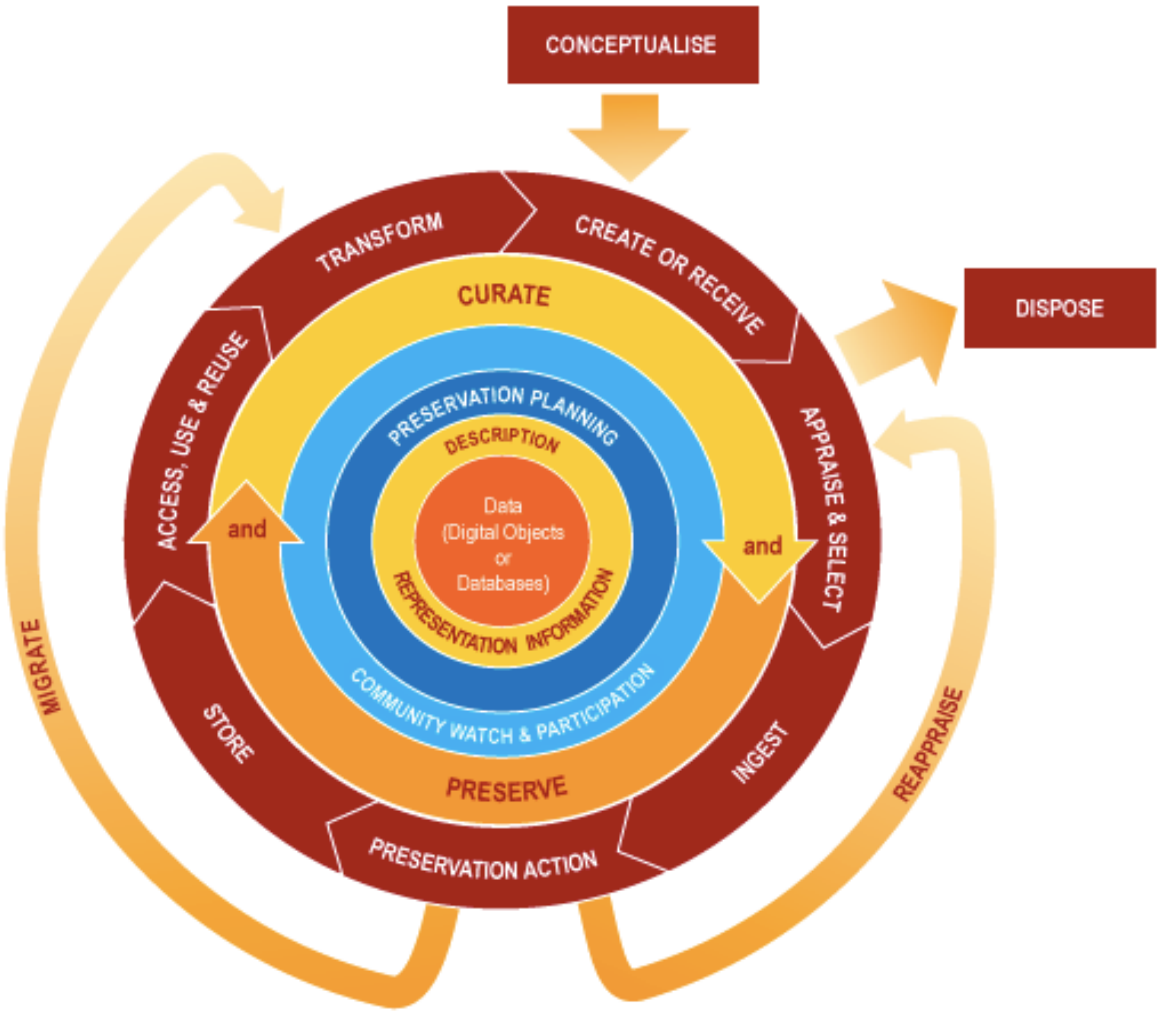

Figura 4: DDC Curation Life Cycle Model
Fuente: Digital Curation Centre (2014)

El modelo expuesto por el DCC proporciona las etapas requeridas para la curaduría y la preservación de los datos desde la conceptualización inicial y puede ser utilizado para planificar las actividades dentro de un determinado proyecto de investigación, organización o consorcio para asegurar que se llevan a cabo todas las etapas necesarias, cada una en la secuencia correcta. Permite definir roles y responsabilidades y construir un marco normativo. Es importante señalar que la

descripción y las acciones de preservación de los elementos del modelo deben ser consideradas en todas las etapas de la actividad. En el centro del modelo se encuentran los datos digitales, que están identificados con objetos o bases de datos simples y complejos. Las relaciones entre las etapas del ciclo de vida presentadas por el modelo señalan los principales niveles de acciones sobre la curaduría de los datos, mientras otros modelos también contemplan las etapas de análisis, es decir, herramientas que sirven como soporte para planificar todo el ciclo de vida de los datos de investigación. Básicamente, todos los modelos de preservación presentan un ciclo en común en relación con su estructura y sus recomendaciones:

- Descripción y representación de la información: creación, recogida, preservación y mantenimiento de los metadatos suficientes para permitir que los datos sean utilizados y reutilizados durante el tiempo necesario.

- Planificación de la preservación: estrategias, políticas y procedimientos para todas las acciones de curaduría.

- Participación y planificación del plan de preservación: la observación de lo que el modelo puede ofrecer para los objetivos de una comunidad específica, es decir, un grupo predeterminado de los interesados en los datos, con el fin de seguir los cambios en sus requisitos para gestión de los datos. Además, en este punto se incluye la participación en la elaboración de normas, herramientas y software relevantes para la preservación de los datos.

La Curaduría y Preservación describe con detalle la mayoría de las acciones del modelo, pero también sirve para representar la ejecución de la gestión de las acciones planificadas y administrativas de apoyo a la curaduría.

Las etapas secuenciales no se ocupan exclusivamente de acciones de análisis, sino que además representan las etapas del ciclo de vida de los datos que debe tener un componente de curaduría. Comienzan con la conceptualización, que señala las etapas de planificación de las actividades de generación y de colección de los datos. Aspectos tales como el método de captura serán informados por distintas consideraciones -el rigor científico del método será especialmente importante-, pero otros asuntos deben tratarse en esta etapa, como qué procedimientos se emplearán para la preservación de los datos, qué presupuesto

se destinará a la curaduría y de qué manera la información importante para la curaduría puede ser automatizada o simplificada.

Las acciones ligadas al proceso de almacenamiento de datos presentan las siguientes etapas fundamentales en el proceso de curaduría:

1) El ciclo de vida comienza con la etapa de *Creación o Recogida de los datos (Create or Receive),* en la que creación se refiere a los datos originales generados y registrados por los investigadores y recogida a los datos obtenidos de otras fuentes preexistentes. En esta etapa las actividades de curaduría deben asegurar que todas las descripciones son suficientes en referencia a los metadatos administrativos, descriptivos, estructurales y técnicos. Pero a medida que diferentes investigadores gestionen los datos inevitablemente deberán describir las diferentes normas que se deben comprobar en referencia a las políticas locales.

2) En la siguiente etapa, *Evaluar y Seleccionar (Appraise and Select)*, los investigadores o curadores evalúan y seleccionan los datos para preservación a largo plazo de acuerdo con el plan de gestión, lo que incluye las políticas o los requisitos legales. En esta etapa algunos datos pueden ser eliminados, lo que puede implicar su transferencia a otro repositorio o su eliminación definitiva. Una vez más, la disposición de los documentos debe ser impulsada por el plan de gestión, las políticas o los requisitos legales.

3) La etapa de *Traspaso (Ingest)* conduce inmediatamente a la fase Acción de Preservación, que implica una serie de actividades: control de calidad, catalogación, clasificación, registro de metadatos semánticos y estructurales, etc. Cualquier dato sin tratamiento debe ser devuelto al investigador para su posterior evaluación. Esto debe dar como resultado una mejora de su calidad (por ejemplo, las correcciones a los procedimientos de transferencia de datos y la adecuación de los metadatos).

4) Una vez que los datos han completado la etapa de Acción de Preservación (Preservation Action), pasan al almacenamiento. Esta etapa tiene que ver principalmente con el compromiso inicial de almacenamiento de los datos, pero ciertas acciones a largo plazo para asegurar que los datos permanecen seguros también pueden estar asociadas con ella: el mantenimiento del hardware de almacenamiento, la realización de copias de seguridad, etc.

5) La etapa siguiente es una secuencia al proceso de Almacenamiento (Store) de la información: se trata de guardar la información siguiendo los estándares establecidos para esos efectos. Para esta etapa es necesario conocer las políticas del repositorio para evitar que puedan afectar al almacenamiento de datos a largo plazo, hay que saber, por ejemplo, cuáles son los formatos más adecuados.

6) Una vez que los datos han sido almacenados de manera segura, entran en el periodo de *Acceso, Uso y Reutilización (Access, Use and Reuse).* Las acciones de curaduría asociadas a esta etapa se centran en mantener, por ejemplo, el control y la adecuación de los metadatos mediante interfaces personalizadas de búsqueda. Aparte de las actividades de preservación en curso, el histórico de los datos archivados se detiene en ese punto; pero diversas circunstancias pueden alterar el ciclo de vida de los datos, como por ejemplo la activación de una acción para migrarlos a un nuevo formato.

7) Por último, la etapa final del ciclo de vida de los datos es la *Transformación (Transformation),* que hace referencia a la necesidad de transformación de los datos a través del tiempo en distintos formatos para evitar la obsolescencia del software. Mediante esta acción los datos vuelven a encontrarse al inicio de su ciclo de vida.

El modelo del ciclo de vida de los documentos "incluye explícitamente las actividades que ocurren fuera del sistema de archivo, es decir, que describe la curaduría en lugar de sólo el archivo o la preservación" (Harvey, 2010, pág. 33) y se utiliza para modelar las actividades de curaduría digital en diversos entornos, como los repositorios institucionales, los archivos digitales y la gestión de los documentos electrónicos. Harvey (2010, pág. 34) señala que comprende tres grupos de acciones: acciones de ciclo de vida completo, acciones secuenciales y acciones ocasionales. Lo que aparece en el centro del modelo hace destacar la importancia de lo que está siendo curado. Las acciones del ciclo de vida de forma interna completan cuatro anillos concéntricos: descripción y representación de la información, planificación de la conservación, observaciones de las comunidades del proceso y "curar y preservar”.

La gestión de los datos de investigación incluye todos los aspectos para preservar datos de investigación y abarca todas las disciplinas. Posibilita que el investigador

mantenga el control de los datos durante todo su ciclo, asegurando la preservación para otros investigadores. Además, evita la duplicación innecesaria de trabajo y favorece el cumplimento de las expectativas y exigencias de las agencias de financiación.

Algunos proyectos pueden utilizar solamente una parte del ciclo de vida; por ejemplo, un proyecto que involucra análisis de imágenes del hielo en la Antártida podría centrarse en integrar y analizar imágenes durante un determinado período, mientras que otro centrado en la recolección de datos primarios y el análisis podría eludir los nuevos descubrimientos con la integración y el análisis de nuevos elementos.

Un investigador o un equipo de científicos se dedican con frecuencia a todos los aspectos del ciclo de vida de los datos, tanto en su faceta de creadores como en la de usuarios. Incluso pueden crear también nuevos datos en el proceso de recogida, integración, análisis y síntesis de los ya existentes. Una adecuada gestión de los datos de investigación también es una manera de mantener la integridad de su investigación, y tiene requisitos fundamentales. En el capítulo siguiente se presentan las principales prácticas para la preparación y gestión de datos.

2.5 Plan de gestión de datos

Un plan de gestión de datos es un documento formal que describe todo el ciclo de vida de los datos desde su recogida hasta la documentación completa del proceso de investigación y registra las decisiones tomadas en relación con estándares de metadatos, formatos, bases de datos, métodos, seguridad y períodos de almacenamiento, así como los costes asociados con la gestión de los datos.

La mayoría de los investigadores recogen datos con algún tipo de plan preconcebido, pero a menudo son incompletos o están documentados inadecuadamente. El plan de gestión de datos requiere una secuencia documentada de acciones destinadas a identificar, asegurar recursos, recopilar, mantener, proteger y utilizar los archivos de datos. Esto incluye la obtención de financiación y la identificación de los recursos técnicos y de personal para el completo ciclo de gestión de los datos. Una vez que se determinen las necesidades

de uso de los datos, el siguiente paso debe ser la adopción de un sistema para almacenar y manipular los que pueden ser identificados y desarrollados.

El alcance y la cantidad de detalles en un plan de gestión de datos dependen del proyecto en sí y del público para el que se está creando. En general, estos planes requieren una descripción del proyecto y de los datos que se generan o utilizan, de los formatos y estándares de metadatos que se emplearán para almacenarlos y organizarlos, de dónde y cómo se almacenan, tanto a corto como a largo plazo, y de las disposiciones de acceso y los requisitos legales que se adhieren a ellos. Los organismos de financiación quieren saber qué han pensado los investigadores sobre la planificación de sus datos digitales y físicos y potencialmente accesibles a un público más amplio.

Aunque los planes sean documentos vivos y pueden sufrir cambios durante un proyecto, todos los requisitos de software, hardware y personal para llevar a cabo el plan de gestión de datos deben estar preparados si se recibe la financiación. Para hacer un registro simplificado del método de trabajo, sea en equipo o de forma individual, se puede utilizar la aplicación DMPTool (2015) como un asistente donde se preparan los planes de gestión de datos mediante un proceso de redacción, e incluso se puede personalizar la herramienta, que es robusta fuera de la plataforma. La aplicación ofrece preguntas de orientación para elaborar respuestas de acuerdo con los requisitos proporcionados por los financiadores y no es necesario ser un centro asociado para utilizarla.

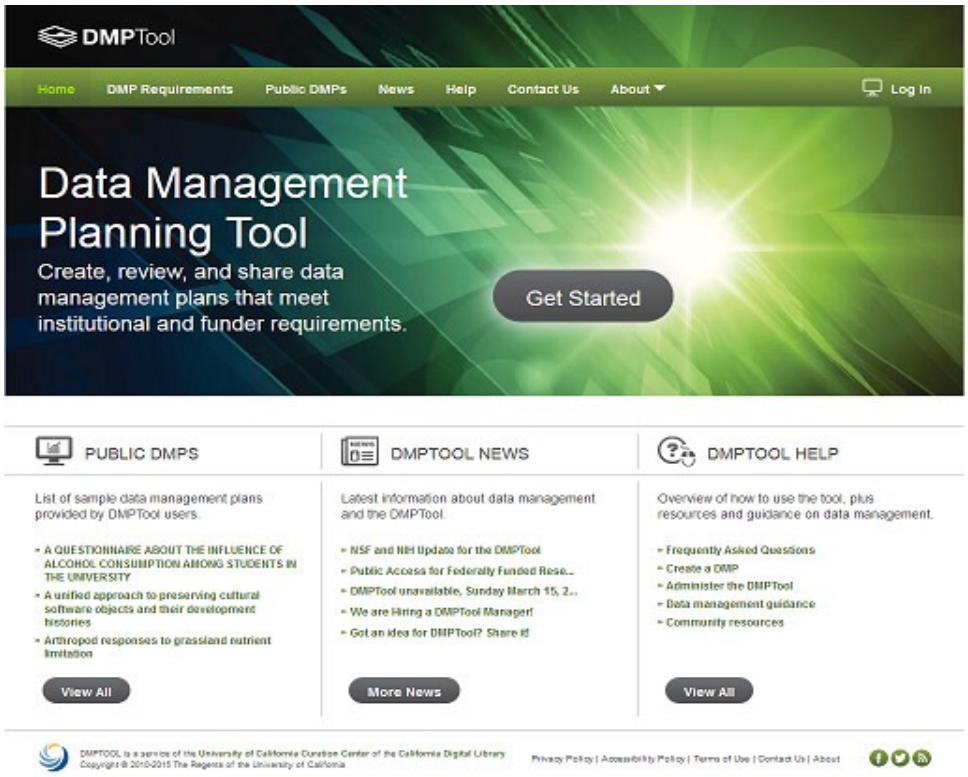

Figura 5: Data Management Planning Tool
Fuente: Digital Curation Centre (2015)

Aunque DMPTool proporciona orientación para los financiadores, también puede ser utilizada por cualquier persona interesada en crear planes de gestión de datos. El objetivo es documentar cómo se producen o se recopilan los datos durante un proceso de investigación y tras su finalización, establecer la forma en que serán descritos, compartidos y conservados. De esta manera todo el proceso queda formalizado en un documento único, prospectivo y descriptivo, así como la evolución de un conjunto de elementos y la información previamente dispersa entre diversos actores y documentos, ofreciendo toda la información necesaria para la supervisión de los proyectos y los resultados.

La documentación del ciclo de vida de los datos posibilita que sean medidos y gestionados desde el momento en que el investigador decide recogerlos o usarlos hasta que se vuelvan obsoletos o ya no sean necesarios. Además, al igual que cualquier otro activo, no puede justificarse o permitirse la adquisición de datos que no sean necesarios. Deben ser adquiridos y mantenidos solo para satisfacer una necesidad científica y han de ser gestionados con el establecimiento de normas y procedimientos para su adecuada presentación.

El DCC proporciona una lista de verificación para un Plan de gestión de datos que reúne información sobre los requisitos de los financiadores y las mejores prácticas en la planificación de la gestión de datos. La tabla 7 presenta los principales temas y preguntas que el investigador necesita responder para cubrir un plan de gestión de datos según DCC.

DCC Lista de verificación	**DCC Guía y cuestiones a considerar**
Datos administrativos	
ID (Identificación)	Una identificación pertinente como determinada por el patrocinador y/o institución.
Patrocinador	Estado del patrocinador de la investigación si fuese relevante
Número de referencia de la subvención	Introducir el número de referencia de la subvención si procede [SOLO CONCESION-POSTERIOR DMPs]
Nombre del proyecto	Si se está solicitando financiación, enuncie el nombre exactamente como en la propuesta de financiación.
Descripción del proyecto	**Temas a considerar:** - ¿Cual en la naturaleza de su proyecto de investigación? - ¿Qué temas de investigación está abordando? - ¿Para qué propósito están siendo recopilados o creados los datos? **Guía:** Resuma brevemente el tipo de estudio (o estudios) para ayudar a otros a comprender el propósito para el cual los datos están siendo recopilados o creados.
PI / Investigador	Nombre del Investigador/es Principal (PI) o Mayor Investigador/es en el proyecto.
PI / Identificador del Investigador	E.g : ORCID http://orcid.org/
Datos de Contacto del Proyecto	Nombre (si diferente al anterior), teléfono y datos de contacto de correo electrónico
Fecha de la Primera Versión	Fecha en que la primera versión del DMP fue completada
Fecha de la última actualización	Fecha del último cambio de la DMP
Políticas Relacionadas	**Cuestiones a considerar:** - ¿Hay algún procedimiento existente en el que basara su enfoque? - ¿Tiene su departamento/grupo directrices de gestión de datos? - ¿Tiene su institución una protección de datos o política de seguridad que usted seguirá? - ¿Tiene su institución una política de Gestión de Investigación de Datos (RDM)? - ¿Tiene su patrocinador una política de Gestión de Investigación de Datos? - ¿Hay algunas normas formales que usted adoptara? **Guía:** Enumere cualquier otra política de gestión de datos, intercambio de datos y seguridad de datos relevante, sea de patrocinador, institución o grupo. Alguna de la información que dé en el resto del DMP será determinada por el contenido de otras políticas. Si es así, señale/enlace aquí hacia ellas.
Recopilación de datos	

Que datos recopilará o creará?	**Cuestiones a considerar:** - ¿Qué tipo, formato y volumen de datos? - Los formatos y software escogidos, ¿permiten compartir y un acceso a los datos a largo plazo? – ¿Hay datos existentes que pueda reutilizar? **Guía:** Dé una breve descripción de los datos, incluyendo cualquier dato existente o fuentes de terceros que serán utilizadas, señalando en cada caso su contenido, tipo y cobertura. Describa y justifique su opción de formato y considere las implicaciones del formato y el volumen de datos en término de almacenaje, reserva y acceso.
¿Cómo serán recopilados o creados los datos?	**Cuestiones a considerar:** - ¿Que estándares o metodologías utilizará? - ¿Cómo estructurará y nombrará sus carpetas y archivos? - ¿Cómo manejara el versionado? - ¿Que procesos de garantía de calidad adoptará? **Guía:** Describa como serán recopilados/creados los datos y que comunidad de normativas (estándares) de datos (si los hubiese) serán utilizados. Considere como serán organizados los datos durante el proyecto, mencionando, por ejemplo, los convenios de denominación, control de versiones y estructuras de carpetas. Explique cómo se controlara y documentara la consistencia y la calidad de la recopilación de datos. Esto puede incluir procesos tales como la calibración, repetir pruebas o mediciones, captura o grabado de datos estandarizados, validación de entrada de datos, revisión por pares de datos o representación con vocabularios controlados.
Documentación y Metadatos	
¿Qué documentación y metadatos acompañarán los datos?	**Cuestiones a considerar:** - ¿Qué información es necesaria para que los datos sean leídos e interpretados en el futuro? - ¿Cómo capturará/creara esta documentación y metadatos? - ¿Qué normativa de metadatos utilizará y por qué? **Guía:** Describa los tipos de documentación que acompañará a los datos para ayudar a usuarios secundarios a comprenderlos y reutilizarlos. Esto debería incluir al menos los datos básicos que ayuden a la gente a encontrar los datos, incluyendo quien los creo o contribuyo con los datos, su título, fecha de creación y bajo qué condiciones se puede acceder. La documentación también puede incluir detalles sobre la metodología utilizada, información analítica y de procedimiento, definiciones de variables, vocabularios, unidades de medida, cualquier hipótesis formulada, así como el formato y tipo de archivo de los datos. Tenga en cuenta cómo va a capturar esta información y donde será grabada. Siempre que sea posible deberá identificar y utilizar las normas existentes en la comunidad.
Ética y Cumplimiento Legal	
¿Cómo manejará cualquier asunto ético?	**Cuestiones a considerar:** - ¿Ha conseguido consentimiento para la preservación e intercambio de datos? - ¿Cómo protegerá la identidad de los participantes si fuese requerido? E.g.: vía anonimización - ¿Cómo se manejarán los datos sensibles para asegurar que se almacenen y transfieran de forma segura? **Guía:** Las cuestiones éticas afectan a la forma de almacenar los datos, quienes pueden utilizarlos/verlos y por cuánto tiempo se conservan. La gestión de las preocupaciones éticas puede incluir: anonimización de los datos; remisión a los comités de ética departamentales o institucionales; y acuerdos formales de consentimiento. Debe demostrar que está al tanto de cualquier problema y ha planeado en consecuencia. Si está llevando a cabo una investigación que involucra a seres humanos, también debe asegurarse de que se solicita el consentimiento para permitir que los datos sean compartidos y reutilizados.

¿Cómo gestionará copyright y temas de Derechos de Propiedad Intelectual (IPR)?	**Cuestiones a considerar:** - ¿Quién tiene la propiedad de los datos? - ¿Cómo se autorizará la reutilización de los datos? - ¿Hay alguna restricción en la reutilización de los datos de terceros? - ¿Se pospondrá/restringirá el intercambio de datos, e.g. publicar o buscar patentes? **Guía:** Exponga quien será el propietario de los derechos de autor y derechos de propiedad intelectual de cualquier dato que vaya a recopilar o crear, junto con la licencia/s para su uso y reutilización. Para proyectos con múltiples socios, puede valer la pena cubrir la propiedad de los DPI en un acuerdo de consorcio,. Considere cualquier política relevante de derechos de autor o derechos de propiedad intelectual de patrocinadores, institucionales, departamentales o de grupos. Considere también permisos para reutilizar datos de terceros y cualquier restricción necesaria en el intercambio de datos.
Almacenamiento y copias de seguridad	
¿Cómo se almacenarán los datos y se realizaran copias de seguridad durante la investigación?	**Cuestiones a considerar:** - ¿Posee suficiente almacenamiento o necesitara incluir cargos por servicios adicionales? - ¿Cómo se harán las copias de seguridad de los datos? - ¿Quién será responsable de realizar las copias de seguridad y la recuperación? - ¿Cómo serán los datos recuperados en caso de incidente? **Guía:** Exponga con qué frecuencia se realizarán copias de seguridad y en que ubicación. ¿Cuantas copias se realizan? Almacenar datos solamente en ordenadores portátiles, discos duros de ordenador o dispositivos de almacenamiento externo es muy arriesgado. Es preferible el uso de almacenamiento robusto, gestionado, proporcionado por los equipos informáticos universitarios. Del mismo modo, normalmente es mejor utilizar los servicios de copias de seguridad automáticos, proporcionados por servicios informáticos, que depender de los procesos manuales. Si decide utilizar un servicio de terceros, debe asegurarse de que esto no entre en conflicto con las políticas de ningún patrocinador, institución, departamento o grupo, por ejemplo en términos de jurisdicciones legales en los que los datos son contenidos o la protección de datos sensibles.
¿Cómo gestionará el acceso y la seguridad?	**Cuestiones a considerar:** - ¿Cuáles son los riesgos de la seguridad de datos y como será esto gestionado? - ¿Cómo controlara el acceso para mantener a los datos seguros? - ¿Cómo se asegurará de que los colaboradores puedan acceder a sus datos de forma segura? - Si se crea o recopilan datos en el terreno ¿cómo va a asegurar su transferencia segura en sus principales sistemas seguros? **Guía:** Si sus datos son confidenciales (por ejemplo, datos personales que aún no están en dominio público, información confidencial o secretos comerciales), deberá describir cualquier medida de seguridad apropiada y mencionar cualquier normativa formal con la que va a cumplir e.g. ISO 27001.
Selección y preservación	

¿Qué datos deberán ser conservados, compartidos y/o preservados?	**Cuestiones a considerar:** - ¿Que datos deben ser conservados/destruidos para fines contractuales, legales o regulatorios? - ¿Cómo decidirá que otros datos conservar? - ¿Cuáles son las investigaciones previsibles utilizadas para los datos? - ¿Por cuánto tiempo los datos serán conservados y preservados? **Guía:** Tenga en cuenta como los datos pueden ser reutilizados, por ejemplo, para validar los resultados de la investigación, realizar nuevos estudios o para la enseñanza. Decida qué datos guardar y por cuanto tiempo. Esto podría basarse en cualquier obligación de conservar determinados datos, el potencial valor de reutilización, lo que es económicamente viable de mantener, y cualquier esfuerzo adicional requerido para preparar los datos para el intercambio y la preservación. Recuerde tener en cuenta cualquier esfuerzo adicional necesario para preparar los datos para el intercambio y la preservación, como el cambio de formatos de archivo.
¿Cuál es el plan de conservación a largo plazo para el conjunto de datos?	**Cuestiones a considerar:** -¿Donde, por ejemplo en que almacenamiento o archivo, los datos serán contenidos? -¿Que coste, si lo hubiese, cobrara su almacenamiento o archivo seleccionado? -¿Ha calculado el coste en tiempo y esfuerzo, de preparar los datos para intercambio/preservación? **Guía:** Considere como los conjuntos de datos que tienen valor a largo plazo serán conservados y curados más allá de la vida de la subvención. También describa los planes para preparar y documentar datos para compartir y archivar. Si usted no propone utilizar un almacenamiento establecido, el plan de gestión de datos deberá demostrar que los recursos y sistemas estarán en posición de permitir que los datos sean curados efectivamente, más allá de la vida de la subvención.
Intercambio de datos	
¿Cómo compartirá los datos?	**Cuestiones a considerar:** - ¿Cómo descubrirán sus datos los usuarios potenciales? - ¿Con quién compartirá los datos y bajo qué condiciones? - ¿Va a compartir los datos a través de un almacenamiento, gestionara peticiones directamente o utilizara otro mecanismo? - ¿Cuándo hará disponible los datos? - ¿Intentara obtener un identificador constante para sus datos? **Guía:** Considere donde, como y a quién deben ponerse a disposición los datos con valor reconocido a largo plazo. Los métodos utilizados para compartir los datos dependerán de una serie de factores tales como el tipo, el tamaño, la complejidad y la sensibilidad de los datos. De ser posible mencione ejemplos anteriores para mostrar un historial de uso eficaz de intercambio de datos. Considere como la gente puede reconocer la reutilización de sus datos.
¿Hay algunas restricciones en el intercambio de datos requerido?	**Cuestiones a considerar:** - ¿Qué medidas va a tomar para superar o minimizar las restricciones? - ¿Por cuánto tiempo necesita el uso exclusivo de los datos y por qué? - ¿Se requerirá un acuerdo de intercambio de datos (o equivalente)? **Guía:** Detalle cualquier dificultad prevista para el intercambio de datos con el valor reconocido a largo plazo, junto con las causas y las posibles medidas para superarlo. Las restricciones pueden ser debidas, por ejemplo, a la confidencialidad, la falta de acuerdos de consentimiento o derechos de propiedad intelectual. Considere si un acuerdo de no divulgación daría suficiente protección para los datos confidenciales.
Responsabilidades y Recursos	

¿Quién será responsable de la gestión de datos?	**Cuestiones a considerar:** - ¿Quién es responsable de implementar el DMP, y de asegurar que se examina y revisa? - ¿Quién será responsable de cada actividad de gestión de datos? - ¿Cómo serán divididas las responsabilidades entre los sitios asociados en proyectos de investigación colaborativos? - ¿Serán la propiedad de datos y las responsabilidades de RDM a ser parte de cualquier acuerdo de consorcio o contrato acordado entre los socios? **Guía:** Detalle las funciones y responsabilidades de todas las actividades, por ejemplo, captura de datos, producción de metadatos, calidad de datos, almacenamiento y copias de seguridad, archivo de datos e intercambio de datos. Considere quien será responsable de asegurar que las políticas relevantes serán respetadas. Las personas deben ser nombradas siempre que sea posible..
¿Qué recursos necesitara para entregar su plan?	**Cuestiones a considerar:** - ¿Se requiere experiencia especializada adicional (o formación para el personal existente)? - ¿Necesita hardware o software adicional o excepcional al existente en el suministro institucional? - ¿Se aplicarán cargos por el almacenamiento de datos? **Guía**: Considere cuidadosamente cualquier recurso necesario para entregar el plan, por ejemplo, software, hardware, expertos técnicos, etc. Cuando se necesiten recursos dedicados, esto debe especificarse y justificarse.

Tabla 7: Checklist para el manejo de datos de la DCC
Fuente: Digital Curation Centre (2015)

El DCC también dispone de herramientas como DMPonline, para la planificación de la gestión de datos, que proporciona orientación a medida y ejemplos para ayudar a los investigadores. El programa hace algunas preguntas al inicio para determinar la plantilla apropiada para cada necesidad y a continuación proporciona una guía para ayudar a interpretar y responder las preguntas del financiador.

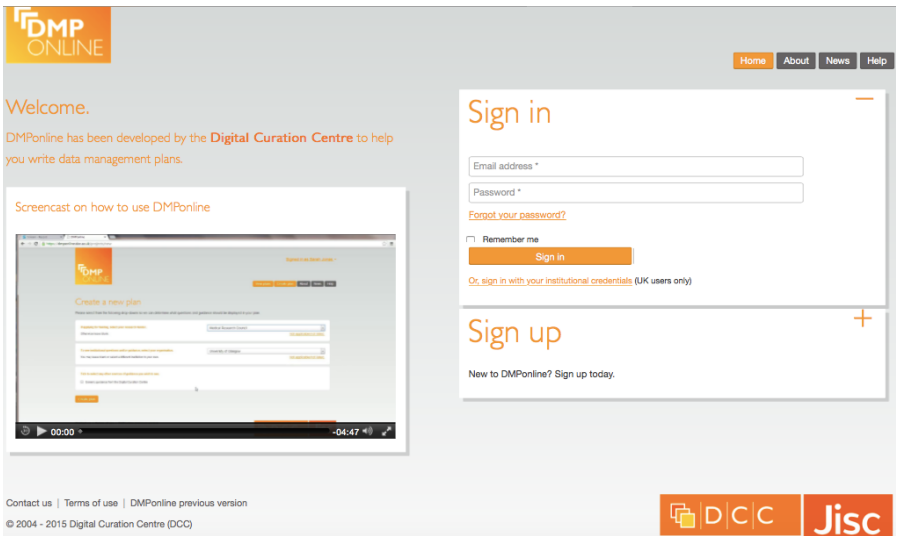

Figura 6: DMPonline
Fuente: Digital Curation Centre (2015)

El establecimiento de planes de gestión de datos es un requisito cada vez más demandado en las convocatorias de proyectos de financiación con fondos públicos, sobre todo en Europa. En muchos casos, las convocatorias requieren los detalles de cómo serán almacenados los datos de investigación durante todo el ciclo de vida, es decir, durante el proyecto y después de su finalización.

En relación con la planificación de los datos de investigación en acceso abierto, por medio del Programa Horizon 2020 la Comisión Europea puso en marcha un proyecto piloto llamado Open Research Data Pilot para fomentar y optimizar la gestión y reutilización de los datos de investigación generados por los proyectos que financia. El uso de un plan de gestión de datos es obligatorio para participar en los proyectos piloto, salvo excepciones justificadas. También se pueden utilizar otros modelos de proyectos y se debe presentar una versión inicial del plan de gestión de datos durante los seis meses posteriores a la aceptación del proyecto. Para atender este requisito del Programa Horizon 2020 se puede utilizar el DMP y presentar un documento corto, de una a dos páginas, pero debe ser actualizado después.

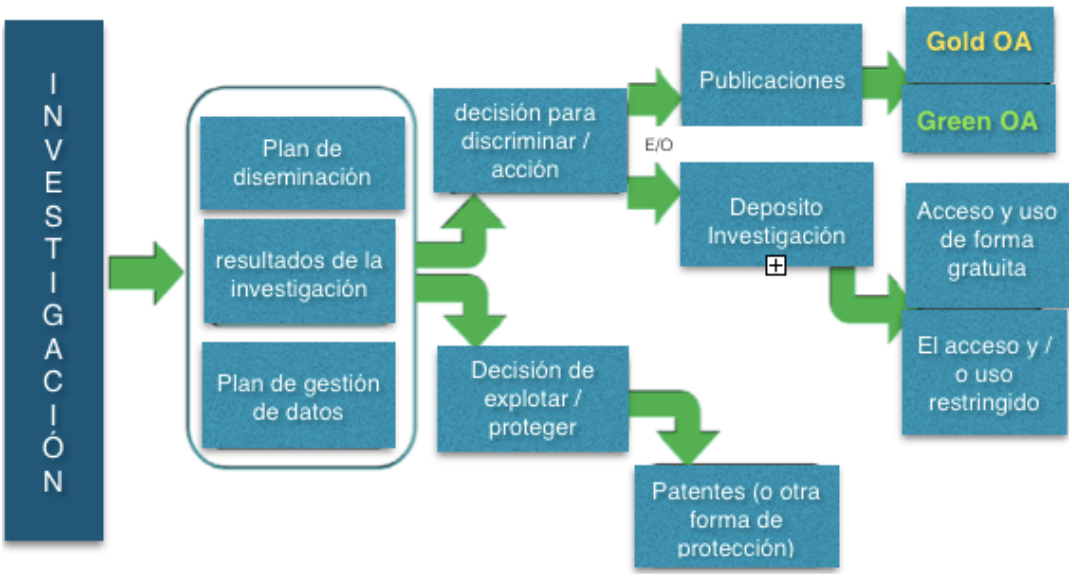

Figura 7: Difusión y explotación de resultados de investigación
Fuente: Programa Horizon 2020

Horizon 2020 representa una oportunidad para poner en práctica no solamente una política, sino también un gran cambio en el desarrollo de proyectos de investigación e innovación de diversas áreas temáticas en el contexto europeo. Se ha implementado un proyecto piloto sobre el acceso abierto a los datos de

investigación y se requieren proyectos participantes para desarrollar un Plan de Gestión de Datos. Por ejemplo, todas las propuestas de proyectos presentadas a las "acciones de investigación e innovación" o "acciones de innovación" en el Programa Horizon 2020 deben incluir una sección sobre la gestión de datos de investigación. El alcance de este programa aún es pequeño; los proyectos pueden optar por el proyecto piloto por una gran variedad de razones, pero sin duda supone un avance hacia la gestión de datos abiertos. Por primera vez, los investigadores están obligados a considerar la preservación y el acceso a sus datos en el inicio de su proyecto.

Los líderes de los proyectos financiados por Horizon 2020 deben dejar libre acceso a los datos de la investigación producidos o recogidos como parte de estos proyectos. Esta distribución gratuita es parte de un círculo virtuoso para mejorar la calidad de los datos, reducir la duplicación de esfuerzos de investigación, acelerar el progreso científico y contribuir a la lucha contra el fraude científico.

Para atender las condiciones del acceso abierto en el proyecto piloto, los beneficiarios deben aceptar el acuerdo de subvención con el compromiso de depositar algunos de los datos (y los metadatos asociados) producidos durante el proyecto en un repositorio de datos de investigación al que asocien una licencia gratuita para su funcionamiento y reutilización. Las excepciones, cuando los datos obtenidos son de uso privado (como datos personales) se negocian con la Comisión Europea en el momento de redactar el acuerdo de subvención. En estos casos, justo antes de disponer de los datos, los responsables de los proyectos deben desarrollar un plan de gestión de datos, incluyendo la descripción de los que no serán publicados.

Así pues, se debe presentar en los primeros seis meses de la investigación un plan de gestión de datos que pueda ser comprobado, revisado y actualizado durante todo el proceso. También se pueden entregar versiones más elaboradas del plan durante el proyecto (European Commission, 2013, p. 3).

Para lograr su objetivo de apertura de los datos, la Comisión Europea pone a disposición el repositorio de datos abiertos Zenodo. Creado por OpenAIRE y CERN con el apoyo de la Comisión Europea, este repositorio de nueva generación ofrece sus servicios a partir de la iniciativa paneuropea OpenAIRE, que amplía la

vinculación de los resultados de la investigación con la información sobre datasets y financiación en contextos europeos y nacionales.

2.6 Los principales componentes de un plan de gestión de datos

Aunque los planes de gestión de datos puedan adoptar muchas formas, siguiendo al UK Data Archive (2015), deben abordarse ciertos componentes principales en todos ellos: 1) Descripción de los datos y metadatos; 2) Actualizar (Metadatos, Documentación); 3) Organización; 4) Adquisición; 5) Procesamiento; 6) Análisis; 7) Preservación; 8) Publicación; 9) Identificadores; 10) Citación de datos; 11) Copia de seguridad; 12) Ética; 13) Propiedad intelectual; 14) Acceso y reutilización; 15) Almacenamiento a corto plazo de gestión y preservación; 16) Almacenamiento a largo plazo de gestión y preservación; 17) Recursos; 18) Personal; 19) Consideraciones para compartir datos; y 20) Formas de compartir los datos.

2.6.1 Descripción de los datos y metadatos

Los metadatos son esenciales al compartir datos, pues hacen posible que los usuarios puedan valorar sus posibilidades de uso (Chapman, 2005), de modo que no sean utilizados de forma inadvertida para fines inadecuados. Los metadatos facilitan el descubrimiento de datos mediante su inclusión en repositorios de metadatos, como el Global Change Master Directory (GCMD)[6] de la National Aeronautics and Space Administration (NASA). Son esenciales en la creación de un registro de paquete de datos, por el que cualquier dato puede ser rastreado hasta su origen.

La diversidad de los datos no se debe solamente a la amplitud de enfoques de los dominios de una investigación, sino también a las muchas formas en que observaciones individuales, objetos, documentos, textos y muestras se pueden representar como datos. Los campos de estudio difieren por estos y muchos otros factores, tales como los objetivos de sus proyectos de investigación, las formas en que se recogen y analizan los datos y sus opciones de nuevas fuentes o recursos existentes para estos. Los investigadores pueden trabajar con un mismo tema y

[6] http://gcmd.gsfc.nasa.gov/

representarlo de variadas maneras y con diversos medios y propósitos con el uso de distintas rutas de análisis, temporales y espaciales, utilizando teorías, métodos y métricas.

En el proceso de representación es esencial para el diseño de infraestructuras de conocimiento eficaces que los datos de investigación estén organizados con metadatos de manera estructurada y con relaciones adecuadas para cada área, con el uso de mecanismos de clasificación tales como taxonomías, tesauros y ontologías.

Extraer valor del volumen de datos disponibles no significa solamente identificar cuáles son útiles; a eso se añade el hecho de que están poco estructurados, se multiplican rápidamente y se diluyen con la misma velocidad con que fueron almacenados. Esos datos requieren nuevas arquitecturas para gestionarlos y manipularlos, de modo que, para extraer valor de su entorno, primero es necesario identificar las ventajas para los investigadores e invertir en la tecnología necesaria para automatizar el proceso de captura, el procesamiento y el almacenamiento de datos. Después, establecer una estrategia de gestión de datos progresiva que habilite a cada sede de almacenamiento de datos para entenderlos e interpretarlos y proporcione como resultado beneficios tangibles. Los datos pueden consistir en cuantificaciones, medidas e investigaciones, y su importancia está en la capacidad de asociarse dentro de un contexto para convertirse en información.

La información se hace inaccesible cuando faltan un hardware y un software sostenibles, cuando son inciertos el origen y la autenticidad de los datos o cuando se ha perdido la capacidad para identificar dónde están los datos, cuál es su localización. Pero también cuando la organización o el proyecto que los originó dejó de existir. Por eso, registrar las observaciones científicas sin una política de preservación de los datos se debe considerar una falta grave para recuperar los resultados de una investigación.

En algunos casos los metadatos están agrupados en repositorios específicos generados por laboratorios y las decisiones operativas en cuanto a utilizar esquemas de metadatos pueden ser informadas por investigaciones empíricas. La utilización de estos resultados estimula la creación de repositorios de registros de metadatos, ayudando a los agregadores de metadatos a no limitar en su trabajo el

empleo de los esquemas disponibles al uso solamente de un pequeño conjunto de elementos (Blower, et al., 2009). Todavía es frecuente que los investigadores no envíen los datos para su publicación, incluso conociendo el repositorio apropiado, lo que demuestra en muchos casos el escaso interés en divulgar los datos de investigación que sostienen las publicaciones de artículos, papers, etc.

Es probable que la descripción de los datos y metadatos de un proyecto sea la mayor parte de cualquier plan de gestión. De acuerdo con UK Data Archive (2015), los creadores del plan deben considerar informaciones básicas sobre los datos que serán recogidos o producidos. También es necesario obtener información acerca de los archivos, formatos y directorios que se utilizarán durante la recogida de datos, incluyendo las convenciones de nomenclatura de archivos, el software que se utilizará para recopilar datos, archivos y formatos. Otro aspecto muy importante es la información sobre las medidas de control de calidad adoptadas durante la recogida de la muestra, durante el análisis y durante el procesamiento de los datos. Además, es fundamental describir quién será el responsable de la gestión durante el proyecto y después de este y cuáles serán sus responsabilidades. Específicamente sobre los metadatos, es imprescindible considerar los detalles contextuales necesarios para que los datos ganen significado (cómo fueron capturados los metadatos, cuándo y por quién fueron creados o capturados y qué normas fueron utilizadas para crearlos y por qué).

El análisis puede incluir la combinación de datos procedentes de varias fuentes. El acceso a cada conjunto de datos individual puede ser simple, pero la conveniencia de analizar múltiples tipos y de ser capaz de hacer frente a grandes cantidades requiere soporte automatizado, que a su vez demanda que los metadatos apropiados estén disponibles.

Corti et al. (2011) argumentan que, para que la toma de datos sea fácil de utilizar, compartible y de larga duración, se debe asegurar que puedan ser entendidos e interpretados por otros usuarios. Esto requiere una descripción de los datos clara y detallada, con información contextual que ponga de relieve el hecho de que, aunque sean procesos conceptualmente distintos, en la práctica la investigación y la preservación de datos no son fácilmente separables. De ahí que la recogida de datos siga en paralelo con su preservación, describiendo y registrando las transformaciones que experimentan. Como los datos brutos se transforman durante

el proceso de investigación, también ocurren cambios en su organización hacia su forma definitiva de conservación.

2.6.2 Actualizar (Metadatos, Documentación)

A lo largo del proceso de ciclo de vida de los datos, la documentación debe ser actualizada para reflejar las acciones emprendidas en su gestión. Esto incluye la adquisición, el procesamiento y el análisis, pero puede afectar a cualquiera de las fases del ciclo. En este proceso es imprescindible actualizar los metadatos para mantener la calidad de los datos. La distinción clave entre los metadatos y la documentación es que los primeros, en el sentido estándar de "datos sobre datos", describen formalmente varios atributos clave de cada elemento o conjunto de elementos de datos, mientras que la documentación hace referencia a los datos en el contexto de su uso específico en sistemas, aplicaciones y configuraciones. La documentación también incluye materiales auxiliares (por ejemplo, notas de campo) de los que los metadatos se pueden derivar. En el primer sentido es "todo acerca de los datos"; en el segundo se trata de "todo sobre el uso”. Ni todos los proyectos utilizarán todos los aspectos del ciclo de vida de los datos ni todos los emplearán de la misma manera. Algunos pueden no seguir los caminos como se muestra y otros pueden rodear de nuevo en ciertos elementos.

Los investigadores pueden documentar sus datos de acuerdo con diversos estándares de metadatos. Algunos están diseñados con el propósito de documentar el contenido de los archivos; otros, para documentar las características técnicas de los archivos, y otros, para expresar relaciones entre los archivos dentro de un conjunto de datos. Para publicar datos de investigación, el estándar de metadatos DataCite sirve como soporte para establecer una estrategia de metadatos que sea capaz de describir sus datos y de satisfacer las necesidades de su gestión.

A continuación se presentan algunos aspectos generales para el registro de los datos de investigación, independientemente de la disciplina. Como mínimo, esta documentación debe ser almacenada en un archivo "readme.txt", o su equivalente, con los datos en sí. También se puede hacer referencia a un artículo publicado que contenga alguna información de la tabla a seguir:

VISIÓN GENERAL	
Título	Nombre del proyecto conjunto de datos o de la investigación que lo produjo.
Creador/es	Los nombres y las direcciones de las organizaciones o personas que crearon los datos; el formato preferido para los nombres de persona es primero el apellido (por ejemplo, Smith, Jane).
Identificador	Número único utilizado para identificar los datos, incluso si es solo un número interno de referencia del proyecto.
Fecha	Fechas clave asociadas con los datos, incluyendo: inicio del proyecto y fecha de finalización, fecha de lanzamiento, período de tiempo cubierto por los datos y otras fechas asociadas con la vida útil de datos, como el ciclo de mantenimiento o la programación de actualización; el formato preferido es aaaa-mm-dd o aaaa.mm.dd-aaaa.mm.dd.
Derecho de acceso	Dónde y cómo los datos pueden ser accesibles y reutilizables.
Métodologia	Cómo se generaron los datos, listado de equipos y software utilizado (incluyendo modelo y versión de números), fórmulas, algoritmos, protocolos experimentales y otras cosas que uno podría incluir en un cuaderno de laboratorio.
Tratamiento	Cómo se han alterado o procesado los datos (por ejemplo, normalizado).
Fuentes	Citas de datos procedentes de otras fuentes, incluyendo detalles de dónde se celebran los datos de origen y cómo acceder a ellos.
Estructura y organización	Relación entre el conjunto de datos y los subconjuntos. Archivos y ficheros con sus correspondientes nombres y extensiones de archivo. Explicación de los códigos y abreviaturas utilizadas.
Financiador	Las organizaciones o agencias que financiaron la investigación.

DESCRIPCIÓN CONTENIDO	
Sujeto	Palabras clave o frases que describan la materia o el contenido de los datos.
Lugar	Todas las ubicaciones físicas aplicables.
Idioma	Todos los idiomas utilizados en el conjunto de datos.
Lista de variables	Todas las variables en los archivos de datos, en su caso.
Lista de códigos	Explicación de códigos o abreviaturas utilizadas en cualquiera de los nombres de los archivos o las variables en los archivos de datos (por ejemplo, '999 indica un valor perdido en los datos').

ACCESO	
Derechos	Ningún derecho de propiedad intelectual conocido, derechos legales, licencias o restricciones en el uso de los datos.
Acceda a la información	Dónde y cómo pueden acceder a sus datos otros investigadores.

Tabla 6: Descriptores de los metadatos
Fuente: Adaptación de UK Data Archive (2015)

2.6.3 Organización

Es necesario describir los tipos de datos y los formatos de los archivos en que estos se almacenan, se mantienen y se encuentran disponibles. Siempre que sea posible, se deben utilizar formatos no propietarios, especialmente al archivar datos o depositarlos en un centro de datos o repositorio.

Para lograr una buena organización, también es necesario documentar todas las acciones que comprende un plan de gestión de datos, incluyendo la obtención de financiación y la identificación de los recursos técnicos y de personal para el completo ciclo de gestión de datos.

2.6.4 Adquisición

La adquisición implica la recogida o la adición a los archivos o fondos de datos. Hay dos formas principales de recogida. La primera es la captura directa mediante alguna forma de medición, como experimentos de observación, encuestas en laboratorio y en campo, mantenimiento de registros (por ejemplo, llenar formularios o escribir un diario), cámaras, escáneres y sensores. En estos casos, los datos están por lo general liberados de medición, es decir, la intención es generar datos útiles. Por el contrario, los datos de escape se producen de forma inherente por un dispositivo o sistema, ya que que son un subproducto de la función principal y no de la salida principal (Manyika et al. 2011). Los datos de escape constan de los diversos archivos generados por los navegadores web y sus plugins, tales como las cookies, archivos de registro, archivos temporales de Internet y archivos .sol (cookies de Flash). Los datos de escape son llamados así por la forma en que

fluyen hacia fuera y detrás del usuario de la web, de manera similar a la forma en que el escape de los automóviles fluye hacia fuera del coche y detrás del motorista.

En otros casos, los datos de escape son de naturaleza transitoria, es decir, que nunca se examinan ni se procesan y son, simplemente, descartados, ya sea porque son demasiado voluminosos o no estructurados en la naturaleza, o porque son costosos de procesar y almacenar, o no existen técnicas para obtener valor de ellos o son de poco valor estratégico (Zikopoulos et al 2012). Por ejemplo, Manyika et al. (2011) señalan que los profesionales de la salud descartan alrededor de un 90 % de los datos que se generan; por ejemplo, casi todo el vídeo grabado en tiempo real creado durante una cirugía.

Los datos de escape son considerados "datos brutos", en el sentido de que no han sido convertidos o combinados con otros datos. Todavía pueden ser procesados para diferentes niveles de derivación en función del uso previsto. Por ejemplo, el Sistema de Observación de la Tierra de la NASA organiza sus datos en seis niveles que van desde los datos capturados en estado bruto por medio de crecientes grados de procesamiento y análisis (Borgman, 2007).

Los datos de investigación, especialmente los de administraciones públicas, son un gran recurso aún en gran parte inexplorado. Muchos individuos y organizaciones recogen una amplia gama de diferentes tipos de datos para llevar a cabo sus tareas. La administración pública es particularmente importante en este contexto, tanto por la cantidad y la centralidad de la recopilación de datos como por el hecho de que esos datos son -o deberían ser- públicos.

2.6.5 Procesamiento

El procesamiento de los datos denota acciones para organizarlos, integrarlos y extraerlos en un formulario de salida adecuada para su uso posterior. Esto incluye archivos de datos y organización de contenidos, la síntesis de los datos o la integración y las transformaciones de formato, y puede incluir actividades de calibración (de sensores y otro campo e instrumentación de laboratorio). Tanto los datos brutos como los procesados requieren el registro de metadatos para asegurar que los resultados se puedan duplicar. Los métodos de procesamiento deben ser rigurosamente documentados para asegurar la utilidad y la integridad de los datos.

2.6.6 Análisis

El análisis implica acciones y métodos que ayudan a garantizar la calidad de los datos al describir hechos, detectar patrones, desarrollar explicaciones y probar hipótesis. La descripción de los datos (metadatos) es importante para localizarlos, comprenderlos e interpretarlos. Es muy útil durante el proceso de investigación y también es un componente crítico para la difusión de datos y su intercambio con otros investigadores.

2.6.7 Preservación

La preservación implica acciones y procedimientos para mantener los datos durante un período de tiempo e incluye su archivo en un repositorio, de modo que estén bien organizados y documentados para facilitar las interpretaciones de los futuros investigadores. Preservar los datos en los centros de datos o repositorios que son gestionados por entidades de confianza para el acceso a largo plazo es la opción más adecuada para compartir datos de investigación.

El almacenamiento basado en la nube contiene datos en servidores remotos, lo que reduce la carga de los problemas de acceso y de gestión. Sin embargo, hay que tener cuidado con la protección al escoger el destino de los datos, principalmente los de acceso privado al ser almacenados en servidores de terceros. Los repositorios y centros de datos son opciones para los conjuntos de datos disponibles para el público, pero no deben ser vistos como almacenamiento primario durante el proyecto de investigación.

Los datos digitales -formados por bits y bytes- son en muchos aspectos más frágiles que los registrados en papel por una serie de razones. Dependiendo del tipo de medios en los que se almacenan los datos (magnético, óptico, etc.), con el tiempo quedan expuestos a diferentes daños o a la descomposición. Cuando se guardan en un servidor, los procedimientos de copia de seguridad y la planificación de recuperación de pérdidas deben tener en cuenta los procedimientos necesarios de acuerdo con el plan de gestión de los datos. Los medios de almacenamiento *off*

line incluyen discos ópticos tales como discos compactos (CD) y discos de vídeo digitales (DVD).

2.6.8 Publicación

La capacidad de preparar y emitir o difundir información de calidad al público para otras agencias es una parte importante del proceso de la gestión de los datos. Hay que asegurar que sean compartidos, pero con controles de propiedad y de integridad de los datos en sí. El intercambio de los datos también requiere metadatos completos para ser útiles a aquellos que los están recibiendo.

2.6.9 Los identificadores

En la elección de un repositorio para datos de investigación es muy importante encontrar uno que les asigne un identificador permanente. El tipo más común de identificador permanente es el DOI (Digital Object Identifier), pero también se puede utilizar PURL (URL permanente) o ARK (Archival Resource Key). La ventaja de utilizar un identificador permanente es que, mientras los URL a menudo cambian con el tiempo, los permanentes siempre deben apuntar al mismo objeto, incluso con los cambios de URL de un documento. Para eso la variedad de esquemas de identificadores presenta propiedades comunes de esquemas de su funcionamiento:

- Accionable (solo accesibles en navegadores web).
- Único a nivel mundial a través de Internet.
- Permanente durante todo el ciclo de vida de los datos.

Los factores más importantes en el almacenamiento y el intercambio de datos a largo plazo son la estabilidad y la buena administración del identificador. Si se mueve el conjunto de datos (por ejemplo, de un repositorio a otro), es posible redirigir la nueva localización apuntando dónde se encuentra mediante su identificador.

Otro factor de consideración es la elección de la dirección URL de host. Este es el nombre de dominio en el comienzo de una URL (justo después de la "http://") que determina dónde comienza la resolución URL (por ejemplo, ub.edu). Un

identificador que no contiene un nombre de host puede utilizar implícitamente otro conocido como punto de partida para su resolución. Por ejemplo, dx.doi.org es el nombre de host para DOI, por lo que el documento identificado por doi: 10.1000/182 se puede encontrar escribiendo "http://dx.doi.org/10.1000/182" en un navegador web.

Los códigos identificadores a largo plazo tienden a ser opacos (por ejemplo, una cadena de dígitos) y revelan poco o nada sobre la naturaleza del objeto identificado, que también es importante para mantener los metadatos asociados con el objeto. Entre las piezas más importantes de metadatos está la dirección URL de destino, que asegura que el identificador siga siendo recurrible. Sea cual sea el identificador o el plan que se elija, si no se actualiza la dirección URL de destino al moverse los datos, el identificador se romperá.

Muchos repositorios emiten solamente un identificador para conjuntos de datos, incluso en "versiones" (carga de nuevos conjuntos de datos, con el histórico de los cambios a los documentos conservados en el repositorio) son permitidos. Pero si el investigador tiene datos que serán actualizados con el tiempo, es recomendable usar un repositorio que emita un identificador con posibilidad de cambios en su registro. La variante DOI puede reflejar lo que las demás versiones de los datos están citando, haciendo referencias a las versiones anteriores del conjunto de datos.

Los identificadores se aplican a cualquier forma de propiedad intelectual. Se utilizan para identificar textos (libros, capítulos de libros, periódicos, artículos, gráficos, etc.), audio, vídeo, imágenes y software. Proporcionan la infraestructura para conectar a los usuarios con el contenido elaborado por los editores, la gestión de la comunicación entre ellos. Permiten la identificación clara y persistente de cualquier tipo de entidad (física, digital o abstracta) en el entorno de Internet. Como ejemplos de objetos a los que se puede asignar un identificador podemos mencionar un CD de música (estructura física), archivos (marco digital) o grabaciones con actuaciones de música (estructura abstracta).

Los identificadores son importantes porque facilitan el rastreo de citas y discusiones de los datos en la Web; de esta manera hacen posible que los datos sean encontrados aunque la URL cambie, que sean transferidos a otro repositorio si el

suyo se cierra, etc. Admiten no solo la identificación y ubicación, sino también la recuperación, y proporcionan referencias cruzadas. En este caso, esas referencias permiten objetos y van a pasajes determinados de texto en los documentos localizados en otras bases de datos, sitios web y revistas fuera del texto en cuestión, es decir, hacen referencia a una dirección web que tiene información sobre el objeto.

La infraestructura del sistema DOI es definida por la norma ISO 26324, llamada “de la información y documentación”: sistema de identificador de objeto digital, promovido por la Fundación Internacional DOI (IDF). La norma deja claro que el nombre se refiere a un identificador digital de un objeto y no a un identificador de un objeto digital. La sigla DOI es asignada permanentemente a un objeto, proporcionando una red de enlace persistente que se refiere a la información actualizada sobre este objeto, incluso cuando este o la información al respecto se puedan encontrar en Internet.

El sistema se consolidó con la creación de la Fundación Internacional DOI (IDF) y las agencias de registro DOI, entre ellos, CrossRef. Esta agencia opera en el contexto de las publicaciones académicas y científicas y es una de las autoridades competentes para el registro y la asignación de identificadores DOI. Debe mantener el control de calidad de los nombres asignados y evitar conflictos de asignaciones.

En la secuencia tenemos algunos esquemas de identificadores:

• ARCA (Archivo Clave de Recursos): una URL con características adicionales que le permiten pedir metadatos descriptivos y de archivo y reconocer ciertos tipos de relaciones entre los identificadores. Es utilizado por instituciones de la memoria, como bibliotecas, archivos y museos. Se resuelve en "http://www.nt2.net". La resolución depende de la redirección HTTP y puede ser administrado por medio de una API o una interfaz de usuario. No hay cuotas de uso.

• DOI (Digital Object Identifier): un identificador que se convierte en acciones concretas cuando se inscriben en una URL. Es muy popular en la publicación de revistas académicas. Se resuelve en "http://dx.doi.org". La resolución depende de la redirección HTTP y el protocolo de identificación de la manija, y puede ser administrado por medio de una API o una interfaz de usuario. Las cuotas anuales se aplican a cada DOI.

Los DOI funcionan gracias a la colaboración de tres tipos de organizaciones. La primera es la International DOI Foundation, que se encarga de la gestión y promoción de los estándares de la marca.

La segunda, la Corporation for National Research Initiatives, que se encarga de desarrollar y mantener todo el sistema para que los DOI se ejecuten correctamente (gestiona el también llamado “Handle System”).

Finalmente, las terceras son las agencias de registro que permiten a las editoriales conseguir los DOI para sus publicaciones bajo los criterios de la International DOI Foundation.

Estas últimas firman acuerdos con el resto de organizaciones para garantizar que los vínculos DOI hacia los contenidos publicados permanezcan inalterables en el tiempo, incluso bajo situaciones extremas como la desaparición de la editorial.

Existen muchas agencias de registro de DOI: Datacite, mEDRA, AiritiDOI, etc. CrossRef es la más grande de todas, con más de 70 millones de DOI registrados en publicaciones de todo tipo: artículos, libros, actas de congresos, paquetes de datos…

Se trata de una asociación de editoriales científicas sin ánimo de lucro que no solo facilita el registro de los DOI a las editoriales, sino que además ofrece al personal investigador servicios y aplicaciones que tienen como base estos códigos.

• Mango: un identificador que se convierte en acciones concretas cuando se inscribe en una URL. Se resuelve en "http://handle.net". La resolución depende de la redirección HTTP y el protocolo de la manija, y puede ser administrado por medio de una API o una interfaz de usuario. Las cuotas anuales se aplican a cada servidor Mango local.

• InChI (IUPAC International Chemical Identifier): un identificador no recurrible de las sustancias químicas que puede ser utilizado en las fuentes de datos impresos y electrónicos, lo que facilita la vinculación de diversas compilaciones de datos.

• LSID (Life Science Identifier): una especie de urna que identifica los recursos de importancia biológica, incluidos los nombres de las especies, los conceptos, los sucesos y los genes o proteínas o los objetos de datos que codifican información

sobre ellos. Al igual que otros URN, se hace recurrible cuando se incrusta en una URL.

• PURL (Persistent Uniform Resource Locator): una URL que siempre se redirige por medio de un nombre de host (a menudo purl.org). La resolución depende de la redirección HTTP, y puede ser administrado por medio de una API o una interfaz de usuario. No hay cuotas de uso.

• URL (Uniform Resource Locator): la dirección típica de contenido web. Es un tipo de URI (Uniform Resource Identifier) que comienza con "http: //" y consiste en una cadena de caracteres que se utilizan para identificar o nombrar un recurso en Internet. Esta identificación permite la interacción con las representaciones de los recursos en una red, por lo general de la World Wide Web, utilizando el protocolo HTTP. Bien gestionada, una redirección URL puede hacer las URL tan persistentes como cualquier identificador. La resolución depende de la redirección HTTP, y puede ser administrado por medio de una API o una interfaz de usuario. No hay cuotas de uso.

• URN (Uniform Resource Name): un identificador que se convierte en acciones concretas cuando se inscribe en una URL. La resolución depende de la redirección HTTP y el protocolo DDDS, y puede ser administrado por medio de una API o una interfaz de usuario. Un navegador plug-in puede salvarlo de escribir un nombre de host en frente de ella. No hay cuotas de uso.

2.6.10 La citación de datos

Sin embargo, el uso o la reutilización de un conjunto de datos debe contribuir a la promoción profesional de cualquier investigador que participe en su recogida o gestión. Varias iniciativas han comenzado a abordar la citación de datos. Un ejemplo de preservación de los datos es la exigencia por parte de varias revistas científicas de publicar secuencias de genes en GenBank, un repositorio público y abiertamente accesible para depositar los datos antes de que se publique el documento. Los datos en sí mismos son compartidos y se vuelven públicos, con la ventaja de que el científico puede depositar el artículo después de ser publicado, mientras que muchas revistas siguen ahora la política de pedir a los autores que sus datos queden disponibles después de la publicación. Sin embargo, parece que

estas peticiones no se cumplen y que la estrategia GenBank de pedir la inclusión del número de acceso en el trabajo supone una mayor garantía de que los datos se harán públicos.

En la actualidad hay pocas maneras de rastrear las citas de datos, aunque esto cambiará a medida que el intercambio y la citación sean cada vez más habituales. La principal fuente de información sobre citas de datos es la Web de Thomson Reuters Data Citation Index. Esta herramienta solo está disponible mediante suscripción, principalmente por medio de una biblioteca institucional. El rastreo de citas de datos aún es limitado; también está disponible mediante el repositorio CrossRef, por ejemplo. El desarrollo de herramientas para el seguimiento de la citación de los datos indica un aumento de los recursos para el control del número de citas, y la tendencia es que con el tiempo los repositorios que aceptan datos serán más fáciles de conseguir.

Si actualmente es posible compartir y reutilizar los datos originales de investigación, es necesario que haya una manera de citar el conjunto de datos de otra persona cada vez que lo utilice. Al igual que con los artículos, los investigadores deben obtener crédito por sus datos compartidos, por lo que no es una obligación citar los que se utilizan. Los procedimientos para citar datos originales de investigación han sido muy recientemente establecidos; por lo tanto, será difícil encontrar esta información en guías de estilo o políticas para publicación de artículos hasta que la citación de los datos de investigación sea algo más habitual.

Al publicar un artículo basado en la investigación que utilice un conjunto de datos externo, se debe incluir la citación para los datos de referencia. Las citas de datos deben aparecer directamente al lado de cualquier cita de artículos. De hecho, la única diferencia entre citar un conjunto de datos y la citación de un artículo es el formato, no la mecánica de la citación. Los datos son un producto igual a cualquier otro de la investigación, y la citación debe reflejarlo. Esto es, en realidad, el principio de la Declaration of Data Citation Principles difundida por Force11, una comunidad de académicos y proveedores de fondos de investigación que han surgido orgánicamente para ayudar a facilitar el cambio hacia la mejora de la creación y el intercambio de conocimientos (Force11, 2015).

La citación en sí debe incluir al menos la siguiente información (Starr; Gastl, 2011):

- Creador
- Año de publicación
- Título
- Editor
- Identificador

Toda esta información en la citación, como el año de publicación y el título de la base de datos, debe estar disponible desde el repositorio de la organización de los datos. Al creador se le identifica con el autor de un conjunto de datos y puede ser una persona, varias personas o incluso una organización. El editor es el depósito que aloja los datos. El identificador es el DOI u otro identificador permanente de los datos; si el DOI no está disponible, se debe utilizar la dirección URL del conjunto de datos. Los identificadores son en realidad una de las partes más importantes de la cita, como la ayuda del DOI con citas de datos de seguimiento.

El estilo de citas elegido, como APA o un estilo específico, puede tener un formato recomendado para las citas de datos, lo cual es útil siempre y cuando incluya la información mínima que aparece arriba. También es posible encontrar una cita con formato de datos en el depósito junto con el conjunto de datos que se esté reutilizando, la cual puede contener o no la información de referencia necesaria (Mooney y Newton, 2012). En ausencia de la exigencia de un formato de cita de datos, es recomendable utilizar el siguiente:

Creador (Año de Publicación): Título. Editor. Identificador

Este es el formato recomendado por DataCite, un grupo internacional de trabajo para estandarizar la citación de datos (Starr y Gastl, 2011). Aquí tenemos un ejemplo de la citación de datos en este formato:

DAO, Smith-Keune C.; WOLANSKI, Jones C.; JERRY, D. (2015). Datos de: Oceanographic currents and local ecological knowledge indicate, and genetics does not refute, a contemporary pattern of larval dispersal for the ornate spiny lobster, Panulirus ornatus in the South-East Asian archipelago. Plos One. http://dx.doi.org/10.5061/dryad.sp418.

La citación de datos debería, como mínimo, incluir los cinco elementos principales ya mencionados; pero también puede incluir otra información, como (CODATA-ICSTI Data Citation Standards and Practices, 2015):

- Versión
- Serie
- Tipo de recursos
- Fecha de acceso

La versión y la serie de información no estarán disponibles para cada conjunto de datos, pero es útil incluirlas en la cita cuando se tengan. El tipo de recurso es nominalmente "conjunto de datos (Data Set)“, pero también puede ser "imagen", "sonido", "software", "base de datos" o "audiovisual". Toda esta información es útil en una citación de datos, pero no esencial.

En algunos casos es posible ofrecer más detalles de los que proporciona el formato de cita datos. Por ejemplo, cuando solo se utiliza una parte de un gran conjunto de datos o se ha utilizado un conjunto de datos al que continuamente se le añaden datos longitudinales o del clima. En estos casos, lo mejor es utilizar el formato de cita estándar y describir más detalles en el texto de los que describe su investigación, lo cual permite mantener el formato estándar sin dejar de ofrecer información adicional (Kratz, 2013).

El formato de cita correcto es el componente más importante para citar un conjunto de datos, pero hay otras consideraciones que se han de tener en cuenta. La primera es que no siempre el depósito en un repositorio ofrece estatus personal para el autor al publicar su trabajo. Esta advertencia a veces aparece en artículos sobre conjuntos de datos de intercambio (Roche et al, 2014), ya que la práctica es muy nueva. Además, mientras sea posible proporcionar crédito del autor a un colaborador al utilizar sus datos, es necesaria solo una citación de autoría principal. Puede ocurrir que el investigador trabaje en estrecha colaboración con un coautor en todo el proceso de investigación; entonces, al hecho de utilizar los datos similares en la construcción de la investigación publicada se le da crédito por medio de la citación. Ciertamente, es posible que el creador de un conjunto de datos

reutilizados pueda convertirse en un coautor, pero el valor predeterminado es la citación.

Otro punto a tener en cuenta es que a veces es posible citar el conjunto de datos independientemente del artículo publicado correspondiente. Por ejemplo, en todos los conjuntos de datos en el repositorio Dryad figuran dos citas, una para el artículo y otra para los datos, con la expectativa de que ambos sean citados. Sin embargo, no es estrictamente necesario citar un artículo cada vez que un conjunto de datos es citado. Los conjuntos de datos no siempre corresponden a un artículo publicado y a veces se puede utilizar el conjunto de datos independientemente de la descripción de su investigación inicial. Sin embargo, a menudo leer el artículo hace entender mejor los datos, lo que da como resultado una citación tanto del artículo cuanto del conjunto de datos. Al final, es el investigador quien debe enjuiciar si es necesario citar el artículo además de los datos.

2.6.11 Copia de seguridad

Es necesario tomar medidas para proteger los datos de pérdidas accidentales y del acceso no autorizado, pues algunos proyectos trabajan con datos que son más sensibles que otros o tienen determinados requisitos de cumplimiento con respecto a la manipulación de los datos. Esto incluye hacer rutinariamente copias adicionales de archivos de datos que se puedan utilizar para restaurar los datos originales o para la recuperación de instancias anteriores de los datos.

Con el fin de mantener la integridad de los datos almacenados, se deben proteger de la manipulación, la pérdida o el robo, limitando el acceso a los mismos por medio de la configuración IP, donde se puede decidir que se autoriza a los miembros del proyecto para acceder a los datos almacenados y gestionarlos. El robo y la piratería son preocupaciones constantes en relación con los datos electrónicos. De acuerdo con el UK Data Archive (2015), muchos proyectos de investigación incluyen el análisis de la recolección y mantenimiento de los interesados en los datos y otros registros confidenciales. Los costes de reproducción con la restauración o el emplazamiento de los datos robados y la pérdida de tiempo que implica la recuperación en caso de robo ponen de relieve la necesidad de proteger el sistema informático y la integridad de los datos.

La seguridad de los datos implica asegurar que estarán a salvo durante el ciclo de vida del proyecto y que solo personas autorizadas podrán acceder a ellos. En este proceso, el uso ético de los datos es importante en los casos en que pueda haber datos sensibles, es decir, los que incluyen información de identificación personal, los que tratan sobre especies protegidas o en peligro de extinción o los que involucran a las poblaciones humanas nativas. El UK Data Archive (2015) señala las siguientes cuestiones para ayudar a identificar el nivel de seguridad de un proyecto:

- ¿Existen obligaciones éticas y/o legales en cuanto a la privacidad y la protección de los datos?
- ¿Cómo se pueden gestionar los datos para protegerlos contra el robo o la piratería?
- ¿Cómo se almacenan los datos? (prestando especial atención a las medidas de seguridad)
- ¿Quién tendrá acceso a los datos y en qué etapa del proyecto?

2.6.12 Ética

Es importante consultar a consejos de ética y asesores legales de carácter institucional para asegurar que los temas de gobernabilidad de los datos hayan sido analizados previamente:

¿Hay problemas éticos y de privacidad que puedan prohibir el intercambio de algunos o todos los conjuntos de datos?

Si existen estos problemas, ¿cómo van a ser resueltos?

2.6.13 Propiedad intelectual

Un ejemplo de un factor de complicación relacionada con la propiedad intelectual es que los datos no pueden ser propiedad del investigador, ya que es esencialmente una colección de hechos, mientras que la organizacion de datos, sin embargo, sí pueden ser considerados como propiedad.

El CONICYT (2010) dispone de una política de acceso a datos de investigación científica que obliga a sus beneficiarios a adoptar las medidas necesarias para permitir el acceso y uso de los datos de investigación resultantes de iniciativas desarrolladas con financiamiento de la Comisión. Se trata de bases de datos no protegidos por derechos de autor. El CONICYT obliga a garantizar que los datos serán efectivamente dados a conocer al público, pero si se trata de bases de datos que poseen protección autoral, el CONICYT exige que sea el propio titular de los datos quien los ponga a disposición.

Las preguntas que siguen ayudan a definir la situación sobre la protección de autoría de un proyecto:

- ¿Quién posee los derechos de los datos creados?
- ¿La propiedad de los datos está cubierta por los derechos de autor?
- Si el conjunto de datos será cubierto por derechos de autor, ¿quién es el dueño esos derechos?
- Si se utilizan los conjuntos de datos existentes, ¿cuáles son las restricciones de licencia?

2.6.14 Acceso y reutilización

Muchas de las necesidades de financiación existentes para los planes de gestión de datos se centran en el acceso a esos datos. Aunque el intercambio de datos sea altamente recomendable por medio de los financiadores, los editores y las instituciones, no es una práctica común en la mayoría de las disciplinas. Exigir que los investigadores describan sus planes para permitir a otros acceder a sus datos y evaluar si se siguieron los planes supone un gran incentivo para que los investigadores compartan sus datos.

Estas son algunas preguntas sugeridas por el UK Data Archive (2015) para que el investigador las tenga en consideración en la planificación del intercambio de los datos:

- ¿Qué datos se comparten?

- ¿En qué etapa se compartirán los datos (en bruto, procesados, reducidos, analizados)?
- ¿Dónde se comparten los datos? (publicados como material suplementario en un sitio web, publicados en un repositorio, etc.).

La citación adecuada de los datos también es importante. Normalmente los investigadores quieren estar seguros de que recibirán el crédito apropiado para los conjuntos de datos que comparten:

- ¿Cómo debería citar sus datos cuando son utilizados por los demás?
- ¿Cómo van a abordar los identificadores la citación de los datos?

2.6.15 Almacenamiento a corto plazo y gestión

Cualquier plan de gestión de datos debe tener en cuenta tanto a largo como a corto plazo la gestión y el almacenamiento de los datos, aunque las preocupaciones y los problemas para los dos marcos de tiempo sean diferentes. Las consideraciones de gestión a corto plazo incluyen cuestiones sobre seguridad de los datos durante el curso del proyecto de investigación, mientras sea necesario prever el trabajo posterior con los conjuntos de datos. Por eso, de acuerdo con el UK Data Archive (2015), los investigadores deben considerar al menos lo siguiente:

- ¿Cómo se manejará el control de versiones?
- ¿Qué datos deben ser almacenados a corto plazo y en qué formatos?
- ¿Cómo será posible el acceso remoto?

2.6.16 Almacenamiento a largo plazo de gestión y preservación

Los planes de gestión de datos deben aclarar los procedimientos previstos para asegurar que los datos estarán disponibles para su uso y reutilización en el futuro. Esto incluye su almacenamiento a largo plazo en un repositorio, los planes para su gestión continua (por ejemplo, la actualización de los metadatos o la información de contacto para el investigador que recoge los datos) y las tareas de preservación (por ejemplo, asegurar formatos de archivo compatibles).

¿Qué datos se conservarán a largo plazo? No siempre es necesario que se archiven todos. Algunos datos preliminares e intermedios no son útiles y los costos asociados a su preservación no se justifican. Si las políticas para mantenimiento de los datos no están claras desde el principio, ¿cómo se decidirá esto en el transcurso del proyecto? En general, para ser utilizados a largo plazo los datos con valor deben ser seleccionados para la gestión continuada. El valor de los datos de larga duración es difícil de predecir; sin embargo, los investigadores más familiarizados con los datos recopilados probablemente serán capaces de identificar la necesidad de su uso durante un periodo prolongado.

Se deben hacer al menos las siguientes preguntas para discriminar los datos previstos para la preservación a largo plazo:

- ¿Dónde se conservarán los datos? ¿Qué repositorios serán utilizados y cuáles serán las políticas de retención?
- ¿Qué metadatos/documentación se presentarán junto a los conjuntos de datos para que sean reutilizables?
- ¿Quién será el responsable de asegurar la vinculación de los datos por medio de códigos identificadores para el proyecto y para las publicaciones pertinentes?

2.6.17 Recursos

Una de las mayores dificultades en la implementación de un plan de gestión de datos es la falta de los recursos necesarios. La gestión de datos implica costos asociados con el hardware, software, personal y similares. Los financiadores animan a los solicitantes de subvenciones para que incluyan estos gastos en el presupuesto, teniendo que considerar los factores que siguen: personal e infraestructuras.

2.6.18 Personal

- ¿Hay recursos suficientes y experiencia en el equipo de investigación para gestionar, preservar y compartir los datos de manera efectiva?

- ¿Qué conocimientos especializados adicionales (o capacitación para el personal existente) son requeridos? ¿Dónde se encuentra este conocimiento?¿De dónde provienen sus recursos?

2.6.19 Infraestructuras

- ¿Qué hardware y software se necesitan para poner en práctica el plan de gestión de datos?
- ¿Hay infraestructura disponible para su uso? Si la hay, ¿es suficiente para gestionar, almacenar y analizar los datos generados por la investigación? En caso contrario, ¿dónde y cuándo se adquieren estos hardware y software?
- ¿Serán adquiridas licencias para software de gestión de datos y herramientas para uso de los investigadores?

2.6.20 Consideraciones para compartir datos

• Formatos de archivo para el acceso a largo plazo: el formato de archivo en el que se mantiene los datos es un factor primordial para garantizar su uso en el futuro, por lo que es necesario incluir en el Plan de Gestión el hardware y software que serán utilizados.

• Documentación: documentar la investigación y los datos para que otros puedan interpretarlos. Es recomendable empezar a documentar los datos al comienzo del proyecto de investigación y continuar durante todo el proceso.

• Propiedad y Privacidad: asegurarse de haber considerado las implicaciones por compartir datos en materia de derechos de autor y de confidencialidad de los entrevistados, siempre que sea necesario.

2.6.21 Formas de compartir los datos

• Correo electrónico a los solicitantes individuales.
• Publicación en línea por medio de un proyecto o sitio web personal.

• Presentarlos como material suplementario alojado en el sitio web del editor de una revista.

• En un archivo o repositorio abierto.

• Compartir directamente con la comunidad científica por medio de redes de colaboración.

Aunque las tres primeras opciones sean formas válidas para compartir datos, la del repositorio es mucho más eficiente para proporcionar acceso a largo plazo.

2.7 Los repositorios de datos

Los repositorios desempeñan una función vital para la preservación, la integridad y la difusión de los datos de investigación. Una red de repositorios genera conexiones entre comunidades, pues cada vez más la interrelación entre fuentes de datos provenientes de distintas disciplinas encuentra lugar en repositorios específicos o multidisciplinares. Los repositorios asumirán el centro de las actividades de búsqueda para la producción del conocimiento, y llegan justo a tiempo, ya que la cantidad de datos de investigación que "nacen digitalmente" aumenta rápidamente.

Con la reciente disponibilidad de la recolección de datos en tiempo real y los avances en potencia de cálculo y espacio de almacenamiento, la capacidad de los investigadores para recopilar grandes cantidades de datos está aumentando. Las discusiones entre dominios científicos se han centrado en la forma de gestionar estos datos y de maximizar nuestro potencial uso, minimizando a la vez su carga de mantenimiento (National Science Board, 2005). Esfuerzos como el proyecto del genoma humano demuestran una capacidad recién descubierta de colaborar a escala global; sin embargo, estas colaboraciones siguen estando bastante arraigadas en ámbitos científicos.

Aunque cada vez más los investigadores compartan sus datos incluyéndolos como un archivo adjunto a los artículos de revistas, los editores están cambiando las políticas editoriales y están requiriendo el archivo de los datos en repositorios. Así pues, reconocen que los repositorios disponen de las características idóneas para una buena organización, preservación y difusión de los datos de investigación y

que, además, facilitan el cumplimiento de los requisitos de la mayoría de los organismos de financiación.

El alcance de estas actividades refleja flujos de financiación de la infraestructura típica. Sin embargo, la investigación es global, y los investigadores necesitan todos los servicios que puedan contribuir a ello. Por lo tanto, hay una necesidad crítica de unión de repositorios de trabajo en todo el mundo. Existen algunas iniciativas internacionales relevantes, tales como las comunidades de metadatos (por ejemplo, la Iniciativa de Metadatos Dublin Core).

Aunque la arquitectura básica de los repositorios es simple (proveedores de datos recogidos por los proveedores de servicios), la realidad es más compleja. Servicios de autoridad, control de acceso (por ejemplo, depósito) que une artículos en repositorios, preservación, estadísticas de uso, creación automática de metadatos, etc., quedan fuera de este modelo simple, y todo funcionaría mejor si se coordinaran a nivel internacional. Sumado a esto, la arquitectura está evolucionando, un ejemplo notable es la aparición de la Iniciativa de Archivos Abiertos (OAI-ORE).

Estrechamente ligada al desarrollo de estándares en la cita de datos, existe una cada vez mayor concienciación de que es necesario conservar, describir y proporcionar el acceso a los conjuntos de datos de manera correcta. Las actividades relacionadas con ello se agrupan en lo que se denomina *data curation* o curaduría de datos. Para que un conjunto de datos sea citado, antes debe haber sido archivado en un repositorio, preservado en un formato interoperable, descrito adecuadamente por un grupo formal de metadatos conectado al conjunto de datos y puesto a disposición de otros investigadores para su reutilización.

En señal de la creciente importancia científica de los conjuntos de datos, Nature Publishing Group ha lanzado la revista Scientific Data (2015), que funciona con revisión por pares y descripciones detalladas de paquetes de datos. Debido a la importancia del grupo Nature en el escenario global de la comunidad científica, su decisión representa mucho en términos de un nuevo concepto para las publicaciones. El objetivo de Scientific Data es promover la reutilización de los datos que sustentan las investigaciones, introduciendo un nuevo tipo de metadato llamado Data Descriptor. Estos metadatos han sido solicitados por la comunidad

académica, las agencias de financiación, revistas, publishers e indexadores, para que los datos de investigación estén públicamente disponibles, sean citables y reproducibles y proporcionen mecanismos de validación que aseguren la calidad y el cumplimiento de las normas de la comunidad científica.

La preservación de los datos de investigación adoptada por la revista Scientific Data es una secuencia obvia, ya que si está confiando sus datos en un repositorio, el investigador quiere saber qué será de ellos hasta que decida quitarlos. Las revistas de la familia PLOS también pasaron a aceptar solamente artículos que acompañen el depósito de datos. En la actualidad, un número limitado de revistas tiene como requisito el depósito de los datos de investigación que hacen referencia a los artículos que publican, y la tendencia es que cada vez haya más publicaciones que incluyan esta exigencia. Evidentemente, el propósito de las revistas es garantizar seguridad a largo plazo, facilidad en la recuperación y el acceso a la comunidad científica.

En relación con los datos oceanográficos, la Marine-Geo Data Library (2015) es un repositorio de datos y metadatos que proporciona un conjunto de herramientas y servicios para el acceso a los datos de investigación de geociencias marinas. Actualmente proporciona acceso basado en la web abierta a 39.0TB de datos, lo que corresponde a más de 573.000 archivos de datos digitales de más de 2.549 programas de investigación que se remontan a la década de1970. Se accede al repositorio por medio de una interfaz de búsqueda y los datos basados en mapas geoespaciales y es accesible por medio de servicios web que siguen las especificaciones del Open Geospatial Consortium (OGC).

Existen diferentes repositorios que pueden alojar los datos en función de su tipología y su área de investigación. Antes de decidirse por una solución para el almacenamiento de los datos, es recomendable determinar cuáles son los repositorios basados en la disciplina específica para archivar los datos. Re3data (Registry of Research Data Repositories, 2015) es un directorio internacional de repositorios que incluye una descripción detallada de los principales repositorios y que, por tanto, puede ayudar a los investigadores a identificar el más adecuado para sus intereses. En España, un proyecto conjunto de investigadores que pertenecen a seis universidades españolas desarrolló el ODISEA, un inventario internacional de los depósitos que admiten conjuntos de datos de investigación a

escala mundial y que permite buscar y sugerir nuevos bancos de datos (ODISEA, 2015).

Mientras exista consenso en que se necesita un conjunto mínimo de información para describir un conjunto de datos, hay dos maneras en que se puede lograr. Una favorece la cita directa del conjunto de datos tal como reside en un repositorio establecido. Este primer modelo fue adoptado para el conjunto de datos de secuencia de nucleótidos en la formación de GenBank (Cinkosky, 1991) y adaptado para las ciencias marinas (Dodge, 1996) y terrestres (Brase, 2004), antes de ser más extensamente recomendado e implementado en varios depósitos de datos de sujetos específicos y generales como la California Digital Library, DataONE, la red Dataverse, Dryad, ICPSR, Pangaea, NOAA's climatic y centros de datos geofísicos y oceanográficos.[7] Uno de los componentes fundamentales de este modelo es la creación y citación de un identificador que únicamente identifica el grupo de datos que está siendo citado. Este identificador generalmente adopta la forma de un “DOI” asignado por DataCite, aunque también pueden ser utilizados otros identificadores.

Este modelo comienza a ser incorporado en los productos de proveedores comerciales de información científica. En 2012 Thomson Reuters lanzó el Data Citation Index[8], una base de datos de grupos de datos que proporciona formatos de citas sugeridos para cada grupo de datos indexado en la base de datos e intenta generar enlaces de citas hacia artículos indexados en su otra base de datos Web of Science. Más recientemente, Elsevier, en cooperación con DataCite y numerosos depósitos de datos, lanzó un proyecto similar que intenta enlazar documentos disponibles en ScienceDirect con los conjuntos de datos que utilizan o tienen depositados mediante grupos de datos “doi” u otros identificadores únicos[9].

La otra forma consiste en la cita de 'artículos de datos' o la 'publicación de datos’ describiendo el conjunto de datos. En este modelo, se presentan los metadatos necesarios para utilizar un grupo de datos junto con un enlace al mismo, en un

---

[7] http://www.cdlib.org/ ; http://www.dataone.org/ ; http://thedata.org/ ; http://datadryad.org/ ; http://www.icpsr.umich.edu/ ; http://www.pangaea.de/ ; http://www.ncdc.noaa.gov/ ; http://www.ngdc.noaa.gov/ ; http://www.nodc.noaa.gov/.

[8] http://wokinfo.com/products_tools/multidisciplinary/dci/

[9] http://www.elsevier.com/about/content-innovation/database-linking

artículo publicado en una revista científica tradicional o en una publicación especializada en datos. Los artículos de datos se diferencian de las publicaciones más tradicionales en que no se requieren análisis o conclusiones resultantes del conjunto de datos. Los investigadores que deseen citar el grupo de datos deberán citar entonces el artículo más que el conjunto de datos. Este modelo ha sido sugerido en las comunidades de neurociencia (Schutter, 2010) (Gorgolewski, 2013), ciencias genéticas (Peterson, 2010) y bioinformática (Chavan, 2011) e implementado en la comunidad de geociencias por medio de revistas de datos como Earth System Science Data[10] y Geoscience Data Journal[11] y de la publicación de artículos de datos en revistas como Quarterly Journal of the Royal Meteorological Society, Eos, y Oceanography.

En algunas ocasiones los repositorios son muy específicos y pueden carecer de vínculos con las publicaciones y otros conjuntos de datos que les den contexto. Algunos pueden tener paquetes de datos inactivos debido a falta de continuidad de una investigación o por otras razones. Por ello es aconsejable comprobar si el tipo de datos del repositorio en el que estamos interesados se actualiza regularmente.

Para comprender mejor la variedad de repositorios existente, hemos establecido cinco grandes categorías que vamos a describir a continuación incluyendo algunos ejemplos.

2.7.1 Institucionales

Los repositorios institucionales ganaron notoriedad en la década de 2000 con el surgimiento de sistemas de software para su implementación como Fedora y DSpace. Tienen por objeto recopilar, administrar y mantener la producción intelectual de una institución académica o de investigación y son plataformas pensadas para la conservación y difusión de las publicaciones científicas (artículos, tesis, documentos administrativos, etc.) generadas por los miembros de la institución. Facilitan la vía verde al Open Access (OA) al proporcionar una vía para

---

[10] http://www.earth-system-science-data.net

[11] http://onlinelibrary.wiley.com/journal/10.1002/(ISSN)2049-6060

que los investigadores autoarchiven todas sus publicaciones, independientemente de la apertura de la revista original que publicó el artículo.

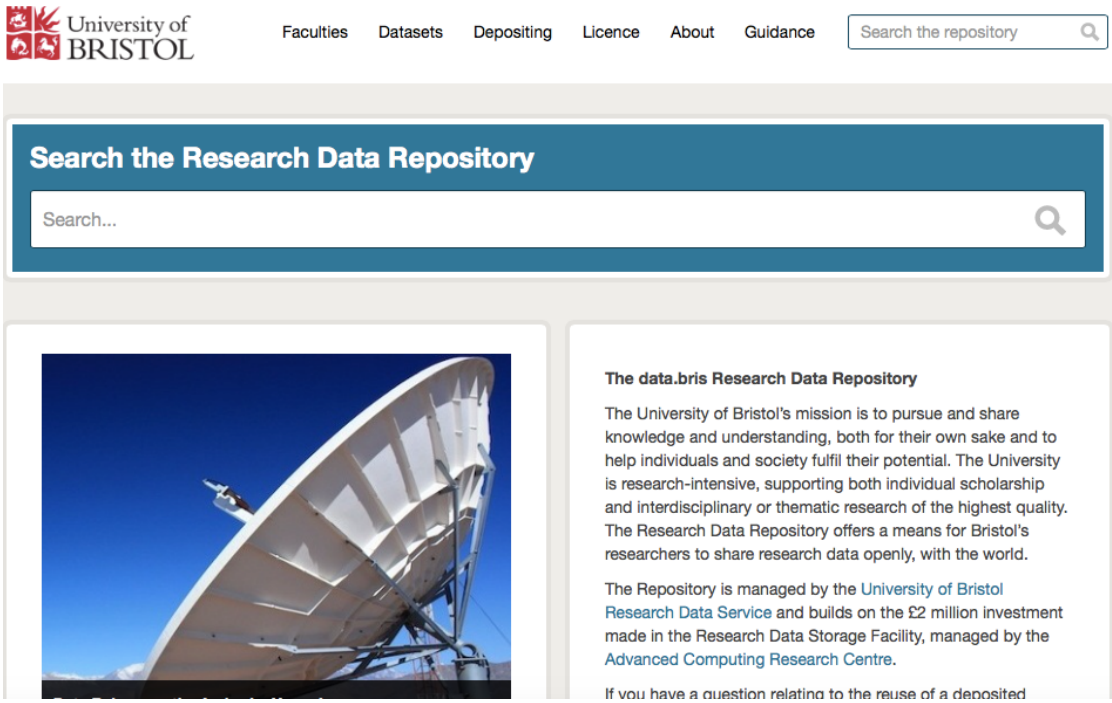

Figura 8: Repositorio institucional
Fuente: University os Bristol (2015)

Con el tiempo, también han permitido el almacenamiento de datos, facilitando la adición de descripciones básicas y complejas de datos, y por lo general emiten identificadores que pueden ser utilizados para citar y encontrar los datos. Algunos RI incluso ofrecen almacenamiento ilimitado de datos y, al estar respaldados por una universidad, normalmente son administrados por los bibliotecarios.

Aunque los RI ofrecen mucha confianza, les falta flexibilidad y control. Muchos de ellos tienen requisitos estrictos para aceptar el archivo de los datos de investigación por medio de formatos muy genéricos, faltan APIs para la interoperabilidad con otros sistemas y muchos solo utilizan un estándar de metadatos muy general como el Dublin Core y no apoyan los campos de metadatos de dominio o de tipo de datos específico y vocabularios controlados.

2.7.2 Temáticos

Los repositorios temáticos son aquellos que incluyen datos de investigación de un campo disciplinario específico. Algunos repositorios temáticos de éxito son ArXiv, Pubmed o Eprints.

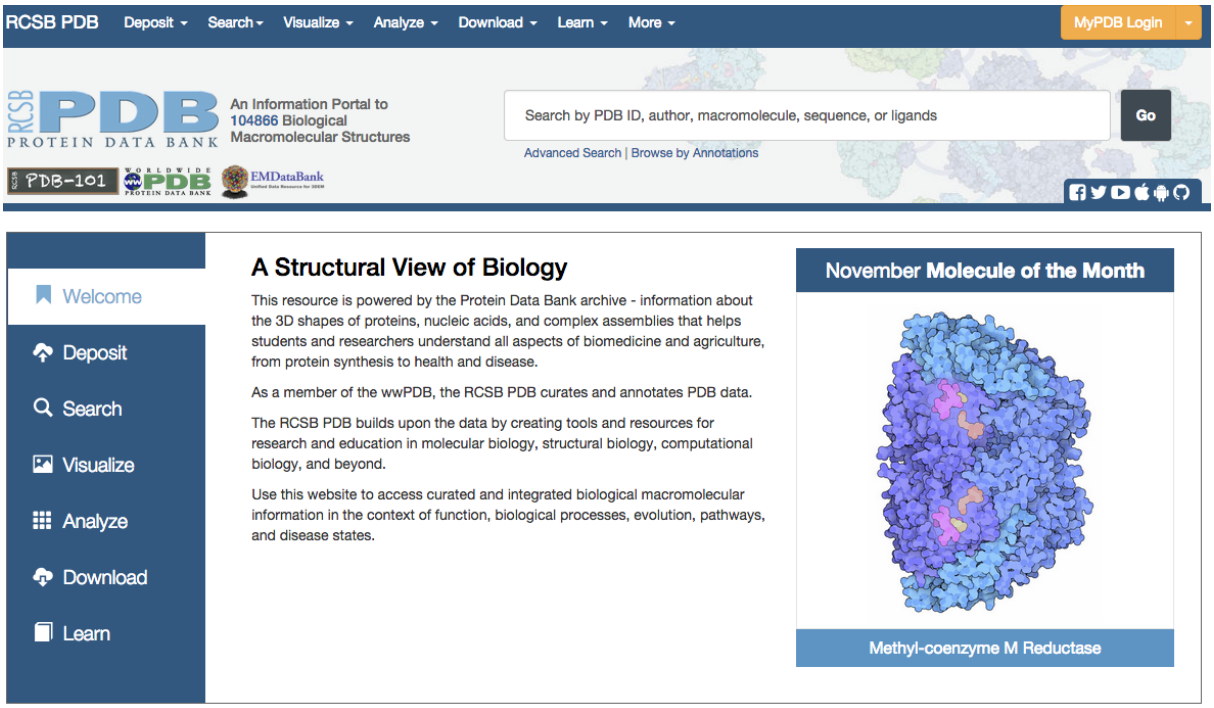

Figura 9: Repositorio Protein Data Bank
Fuente: Protein Data Bank (2015)

Son diversas las disciplinas que disponen de repositorios diseñados específicamente para los tipos de datos de su dominio. Algunos ejemplos son el Protein Data Bank de la Research Collaboratory for Structural Bioinformatics (RCSB), para formas 3D de las proteínas, ácidos nucleicos y conjuntos complejos; el GenBank, para las secuencias de ADN; el EMDataBank, con mapas 3D de microscopía electrónica de densidad, modelos atómicos y metadatos asociados; el eCrystals, para datos cristalográficas de rayos X, y la National Oceanographic Data Center (NODC), para datos oceanográficos.

A menudo, estos repositorios temáticos tienen herramientas analíticas y de descubrimiento disponibles junto con los datos para fomentar su reutilización. Algunos expertos sugieren que los datos deben ser alojados únicamente en repositorios temáticos porque, según ellos, permiten el uso especializado de metadatos y una mayor revisión y validación por expertos en el campo. Sin embargo, no todas las disciplinas tienen repositorios de datos y la propia

especificidad y peculiaridad de muchos datos explica las dificultades para encontrar almacenamiento en los repositorios existentes.

2.7.3 Editoriales

Los repositorios editoriales ofrecen características similares a los institucionales, pero con características especiales para comunidades específicas.

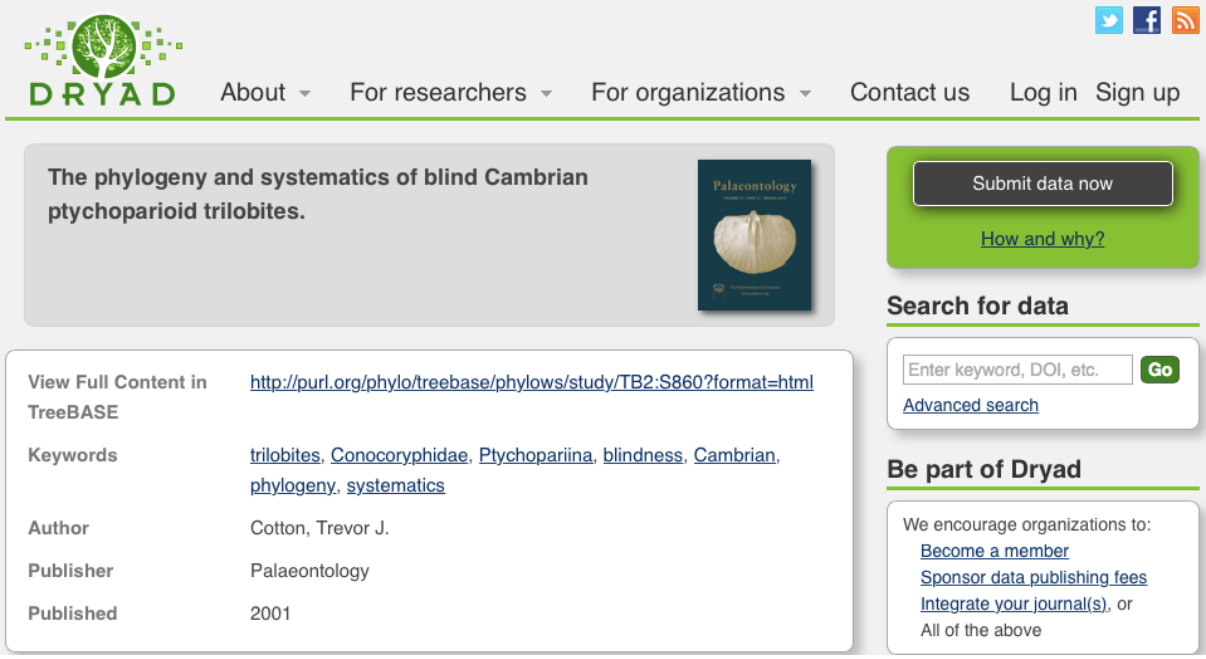

Figura 10: Repositorio Dryad (ejemplo de registro)
Fuente: Dryad (2015)

Dryad es un repositorio digital de datos de investigación científica y médica, de carácter internacional, procedente de revistas científicas revisadas por pares. Funciona como repositorio de diversas disciplinas y también facilita el código DOI, asignado mediante el servicio EZID de la California Digital Library y registrado por DataCite. Cuenta con revistas tanto de acceso abierto como comerciales. Una de sus principales características es la capacidad de albergar cualquier tipo de datos huérfanos. Además, Dryad minimiza la carga de presentación de los artículos, es decir, el repositorio hace una lectura automática de los metadatos proporcionados por las revistas asociadas que proporcionan la información bibliográfica de cada artículo antes de su publicación.

2.7.4 De propósito general

Se trata de repositorios que cualquier investigador puede usar, independientemente de su afiliación institucional, para preservar cualquier tipo de producción académica. Los dos ejemplos más conocidos son Figshare y Zenodo.

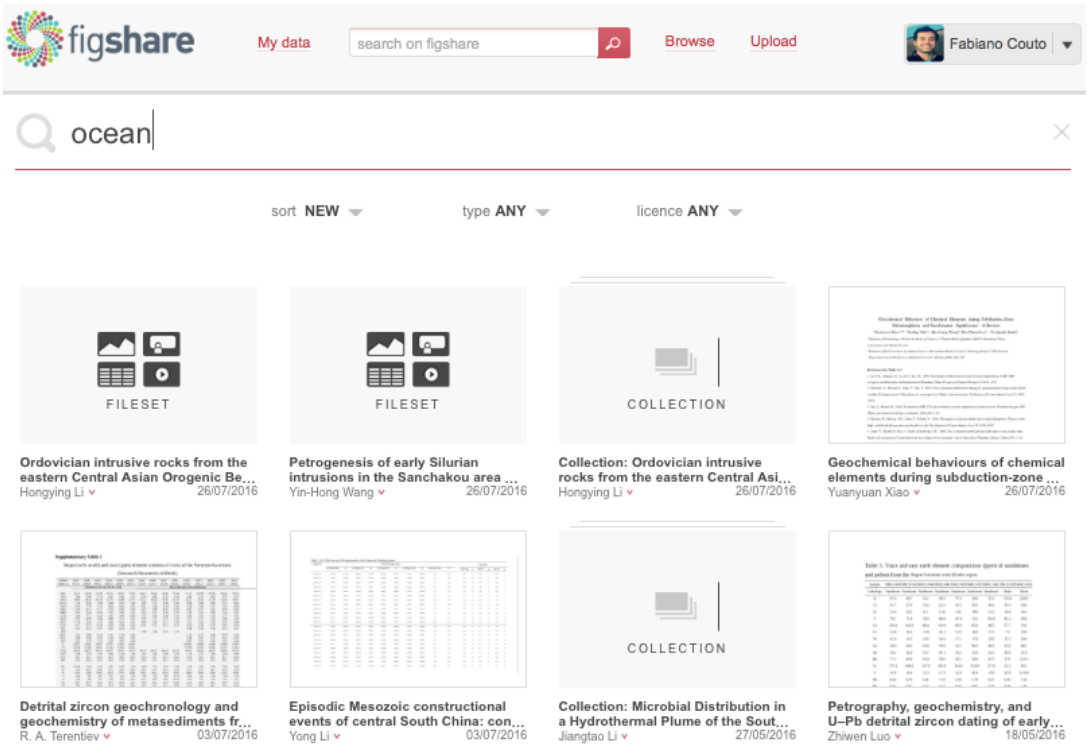

Figura 11: Figshare (página de resultados)
Fuente: Figshare (2015)

Figshare es una plataforma creada por Digital Science que permite compartir y mostrar los resultados de investigaciones multidisciplinares y que está dirigida a investigadores, científicos, proyectos e instituciones. Actualmente está asociada con F1000 Research (un prestigioso repositorio de artículos científicos), colabora con PLOS (la revista científica de acceso abierto más grande del mundo) y también con Plum Analytics (un servicio que cuantifica el impacto de los trabajos de investigación publicados). Todo el material publicado en Fisgshare es identificado con un DOI para facilitar su localización y su cita. En la plataforma podemos localizar: presentaciones, vídeos, pósters, imágenes, datos, artículos, etc. y la preservación de los datos funciona con tecnología CLOCKSS, una organización sin ánimo de lucro que promueve la alianza entre los editores del mundo académico y

las bibliotecas académicas para archivar de un modo sostenible todo el contenido web producido en el ámbito científico.

Los usuarios pueden integrar los datos del repositorio con otros sitios web y blogs copiando y pegando un simple código. Los lectores pueden hacer comentarios sobre los conjuntos de datos y descargar archivos de citación a sus gestores de referencia para su uso posterior. El repositorio también ofrece la posibilidad de publicar resultados negativos o sobre experimentos fallidos para que otros investigadores se ahorren el esfuerzo de tener que pasar ensayos ya realizados, y así no malgasten muchas horas de trabajo en determinados casos.

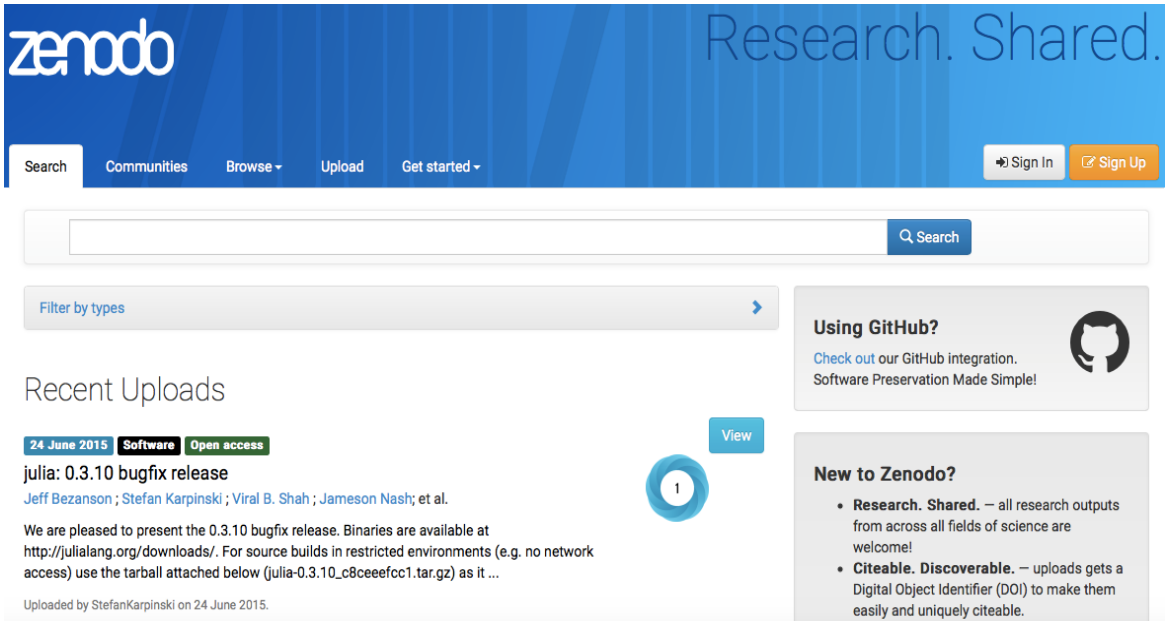

Figura 12: Repositorio Zenodo (página de inicio)
Fuente: Zenodo (2015)

Zenodo es una iniciativa del portal OpenAire que dispone de una infraestructura adecuada para el alojamiento de datasets y otros resultados de investigación de proyectos europeos. Está construido sobre la plataforma Invenio y desarrollado en el CERN, el centro que se ocupa también de la gestión de la enorme cantidad de datos del Gran Colisionador de Hadrones (LHC). Como en el caso de Figshare, el acceso al depósito es libre, asigna DOI y permite conjuntos de datos disponibles en BibTeX, EndNote y otros formatos bibliográficos. Los usuarios pueden añadir metadatos a sus archivos, mucho más detallados que en Figshare. Todos los datos son susceptibles de ser recolectados por otras plataformas mediante el protocolo OAI-PMH. Zenodo impulsa la carga amigable de los datos gracias a la

comunicación con servicios como Mendeley, DropBox, CrossRef o ORCID. También contempla estrategias de preservación digital a largo plazo, permite establecer licencias flexibles para gestionar los derechos y permite a los usuarios crear sus propias colecciones en un espacio propio utilizando metadatos bajo licencia CC0, es decir, dedicadas al dominio público sin restricciones ni solicitud de permisos, excepto para las direcciones de correo electrónico. Además, siempre que esté permitido, otros usuarios Zenodo pueden comentar sus archivos, y una interesante característica es que hace que sea fácil inscribirse con su identificador ORCID o cuenta de GitHub.

2.7.5 Repositorios propios

En algunas ocasiones, los investigadores archivan sus datos científicos en un servidor personal o de sus proyectos situado en la nube. Se pueden encontrar variadas opciones tecnológicas, como la gratuita de Dropbox, y otras comerciales, como las que ofrecen Amazon Cloud Drive o Microsoft Azure. El mantenimiento de los repositorios propios depende de la capacidad del investigador para adaptar las necesidades de sus proyectos, así como para llevar a cabo acciones de backup y replicación adecuadas. En este sentido, se trata de una de las opciones menos recomendables por la falta de garantías en la organización, el mantenimiento y la preservación.

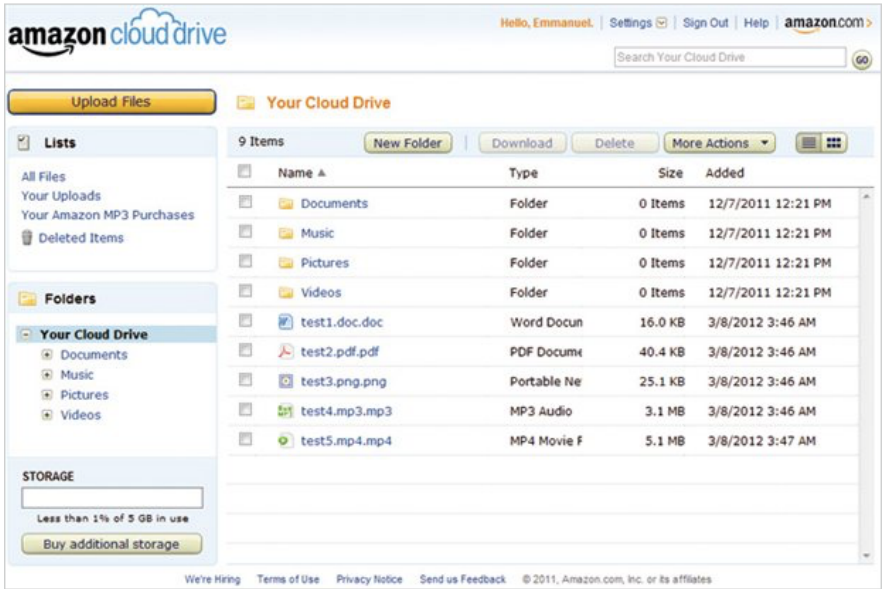

Figura 13: Repositorio Amazon Cloud Drive (cuenta de usuario)
Fuente: Amazon Cloud Drive (2015)

2.7.6 Análisis comparativo

Como hemos visto, los repositorios reúnen los conjuntos de datos que acompañan publicaciones científicas diversas, aunque en algunos casos su función sea archivar solamente los datos de investigación, enlazando los ficheros para las publicaciones que los sostienen. Los datos deben presentarse a repositorios específicos, reconocido por la comunidad siempre que sea posible (temáticos y institucionales), o a repositorios generalistas (Figshare, Zenodo, Dryad, etc) compatibles con la investigación. La selección debe asegurar que cumplen con los requisitos de acceso a datos, la conservación y la estabilidad. Es necesario tener en cuenta, sin embargo, que algunos repositorios sólo pueden aceptar datos financiados por fuentes específicas, o cobran por alojamiento de datos.

Es recomendable que, en los casos que los datos no hayan sido depositados en un repositorio antes de la presentación de manuscritos, los autores envíen sus datos a figshare o el Repositorio digital Dryad durante el proceso de envío. También se puede depositar los datos temporalmente en estos recursos, si el repositorio principal anfitrión no acepta datos en casos específicos, como las investigación aún en desarrollo, por ejemplo.

Sin embargo, los repositorios generalistas puede manejar una amplia variedad de datos, y también pueden ser apropiados para el almacenamiento de los análisis asociados, o datos de control experimental, que permiten a los autores cargar archivos a estos recursos, junto con su manuscrito de datos de descriptores, complementando el registro de datos primarios.

Los repositorios de datos institucionales sólo aceptan los datos generados por los investigadores de la misma institución, mientras que los repositorios específicos del proyecto sólo aceptan los datos de un proyecto específico.

Los repositorios institucionales en instituciones académicas tienen el objetivo de preservar y poner a disposición una parte del trabajo académico de sus estudiantes, profesores y personal. No todos los repositorios tienen la capacidad de aceptar la curaduría de datos.

3 DATOS OCEANOGRÁFICOS

En este capítulo se analiza la situación de los datos oceanográficos y se presentan las bases para la gestión de este tipo de datos desde las políticas y prácticas administrativas y técnicas institucionales.

En común con el capítulo anterior (Los datos de investigación), describiremos la tipología y el ciclo de vida de los datos de investigación, con énfasis en la oceanografía.

3.1 La oceanografía

La oceanografía es la ciencia que tiene como objeto de estudio los océanos y los mares de la tierra y se caracteriza por tener un carácter multidisciplinar. Esto se debe al hecho de que su campo de investigación consta de cuatro disciplinas: la Oceanografía física, que se encarga de los fenómenos físicos que tienen lugar en el océano, como las olas y las corrientes; la oceanografía biológica, la cual realiza investigaciones sobre los seres vivos que habitan en el océano; la oceanografía química, centrada en el análisis de la composición acuática desde el punto de vista químico y la Oceanografía geológica, que investiga el desarrollo geológico que incide sobre la conformación del mar y de la costa. Cada una de ellas comporta una recopilación de datos estadísticos, que permiten describir el componente medioambiental de las áreas geográficas donde se llevarán a cabo las operaciones.

Así pues, la oceanografía es la ciencia que investiga las características de los océanos, mares, ríos, lagos y zonas costeras en todos los aspectos de su descripción física y la interpretación de los fenómenos que se producen en ellos, y su interacción con los continentes y la atmósfera. Además, investiga los animales y las plantas, el medio ambiente y los procesos marinos. Recoge, analiza e interpreta la información sobre las características físicas, químicas, biológicas y geológicas del medios acuático. Analiza la composición del agua de los ríos, lagunas y estuarios y trabaja en proyectos de saneamiento en las zonas costeras.

En oceanografía, muchas variables físicas, como la salinidad, la temperatura y la velocidad de las corrientes se pueden obtener a través de la recolección de datos (Silva, 2006). Por lo tanto, las instituciones ambientales, tales como la National oceanographic and Atmospheric Administration (NOAA) utilizan estos datos para crear modelos numéricos y poner los resultados a disposición del público en Internet. Vale la pena señalar que estos modelos se procesan y, por lo tanto, genera un gran volumen de datos con en el tiempo.

En este sentido, un factor muy importante en el desarrollo de conocimientos relacionados con la corriente oceánica y otros aspectos asociados es el uso de los resultados generados por las simulaciones numéricas. Esto es posible gracias al bajo costo de operación, así como a la rapidez de los ordenadores. Esta configuración posibilita de estudiar, de forma rápida, extensas áreas geográficas y probablemente no podría estar cubierto por un crucero oceanográfico (Cirano et al., 2006), o mediante mediciones con instrumentos oceanográficos.

En relación a los estudios polares, el océano Polar Antártico tiene una de las mayores corrientes oceánicas en la Tierra y es uno de los más rápidos. En algunas partes, como el estrecho de Drake (paso entre el Pacífico y el Atlántico en América del Sur y el norte de la Península Antártica) puede alcanzar una velocidad de 60 kilómetros. Esta corriente se conoce como la corriente Circumpolar Antártica y camina junto a la línea de la convergencia antártica, a 60ºS (el lugar donde se puede obtener alrededor de toda la Antártica por mar).

Existen interacciones entre las corrientes profundas del lecho de los océanos del mundo y tres de las aguas superficiales alrededor de la Antártida. En el Atlántico, por ejemplo, se alternan con agua desde el hundimiento de la costa de la Antártida hasta el fondo del Atlántico, en dirección hacia el norte, mientras que en una profundidad intermedia. Por el contrario, las aguas del Atlántico Norte, pasan entre estas dos corrientes y se unen a la corriente Antárctica, la puerta de África del Sur. Tales movimientos de masas de agua son responsables de diversos procesos. La oceanografía química, biológica y geológica permiten desentrañar una maraña de secretos detrás de la

dinámica de los océanos y su interacción con la corriente antárctica. Estas conexiones indican cómo todo en el planeta está interconectado.

Para llevar a cabo la investigación oceanográfica se utilizan instrumentos capaces de medir propiedades físicas y químicas de la columna de agua, del fondo marino y de las interfaces tierra, agua y aire. Algunos de estos instrumentos llevan décadas empleándose y han evolucionado poco en cuanto al principio de medida de la captura de los datos, pero hay una gran mayoría de instrumentos de nueva generación, siempre en evolución, que recogen una gran cantidad de datos multivariables. Su empleo conjunto hace posible el estudio de los procesos integrados al movimiento de los mares y a la predicción del comportamiento, y además sirve como método para realizar calibraciones cruzadas o ajustes entre los distintos equipos de adquisición de datos.

Los estudios oceanográficos se asientan en datos recopilados directamente del ambiente marino, aunque el uso de la tecnología de percepción remota y la modelación numérica suponen igualmente una fuente de datos oceanográficos. Estos datos asumen conexiones con cada una de las disciplinas de estudio de la Oceanografía y de las disciplinas relacionadas con ésta, tales como la Limnología, la Hidrografía, la Hidrología, la Meteorología y las Ciencias Ambientales, entre otras.

3.2 Los datos oceanográficos

En secuencia al desarrollo de investigaciones, la Oceanografía comprende una recopilación de información *in situ*, que pone en marcha modelos oceanográficos con la asimilación de datos reales, produciendo, a su vez, una mayor precisión de las estimaciones a corto plazo. Por último, consiste en la producción de las previsiones de diversos parámetros en el corto y mediano plazo, por los modelos de carrera con la asimilación de datos adquiridos *in situ*. Esta adquisición se lleva a cabo principalmente a través de sondas de CTD (conductividad, temperatura y profundidad), que permiten la obtención de los valores de la profundidad, la salinidad y la temperatura, entre otros, de

la superficie de la columna de agua a una profundidad de interés operativo (Pacheco y Martinho, 2005) e investigaciones observacionales.

Instrumentos científicos y simuladores generan datos obtenidos por sondas acústicas, sondas de conductividad, temperatura y profundidad, medidores de corriente, entre otros. Estos datos precisan de una infraestructura en red para ser gestionados y que así sean aprovechados en todo su potencial por las comunidades de investigadores (Laaksonen,, Schroeder, Arzberger, Casey, Bowker, Beaulieu, et al., 2004). Además, la proliferación de dispositivos utilizados por la comunidad oceanográfica conectada a Internet (smartphones, tablets, ordenadores, cámaras de vídeo, GPS, etc.) difunden y almacenan diariamente millones de bytes de datos en la red. Esta enorme cantidad de información no se limita al almacenamiento en grandes servidores, sino que también es susceptible de ser procesada y analizada correctamente para obtener el mejor valor posible. En el campo de las investigaciones oceanográficas, la captura, organización y el manejo de estos conjuntos de datos pueden favorecer la toma de decisiones basadas en la recolecta de datos geográficos de forma rápida y con resultados efectivos.

Estos datos son recogidos de las profundidades del océano y también de las zonas costeras, de los estuarios y de los casquetes de hielo polares de la Antártida, teniendo en cuenta los estudios realizados por diferentes sectores relacionados con el medio marino. Para almacenarlos, existen centros de investigación con capacidad para recibir información permanente de diversas fuentes que pueden ser tratadas M2M (machine to machine). Las aplicaciones existentes de software permiten a los investigadores recibir y enviar datos desde sus laboratorios, mientras que las herramientas hoy disponibles no son suficientes para manipular automáticamente la variedad de datos oceanográficos.

Al mismo tiempo que los datos oceanográficos son utilizados con objetivos operacionales o manipulados por científicos para el desarrollo de publicaciones científicas, también son un recurso en sí mismos. Pueden ser reutilizados si han sido debidamente almacenados, sirviendo como recurso renovable para el avance científico y comercial. De acuerdo con la Comisión

Oceanográfica Intergubernamental de la Unesco (2007, p.7), los datos oceanográficos normalmente son recogidos de distintas formas: se instalan sensores, se arrastran redes, se largan instrumentos desde buques, que se dejan a la deriva o se amarran a cables y plataformas; los satélites observan los océanos desde el espacio; y se construyen laboratorios en el fondo del mar. Aun sumado se efectúan mediciones con una gran variedad de fines, mediante personas y sensores con el apoyo de diversas instituciones, entre ellas los poderes públicos, empresas privadas y organizaciones no gubernamentales.

Figura 14: Fuentes de datos oceanográficos
Fuente: elaboración propia (2014)

Los datos oceanográficos son insustituibles teniendo en consideración las variantes espaciales y temporales en el momento de ser recogidos. Un atenuante son los costos para obtención de datos oceanográficos, una vez que “hay una creciente necesidad de datos operacionales en tiempo casi real con fines de predicción del estado del mar” (COI, 2007, p.7).

3.2.1 Datos de las investigaciones polares

Las dificultades a las que se enfrentan los científicos para tener acceso a los datos generados en un lugar de logística tan compleja como la Antártida, puede plantear un desafío de costosas consecuencias para las instituciones que promueven la ciencia. La investigación en el entorno polar es lenta, a pesar de que existan muchos datos recogidos. El correcto almacenamiento de los resultados puede ser una herramienta científica importante, así como la generación de enfoques metodológicos estandarizados para facilitar la organización de metadatos.

A escala mundial, debido a la complejidad inherente de la agrupación y organización de la información generada por la investigación polar, una de las pocas acciones exitosas destinadas a almacenar los datos en la ciencia antártica fue el lanzamiento del Antarctic Master Directory (AMD). Gracias al apoyo de la agencia espacial americana (NASA) y del comité permanente SCAR sobre gestión de datos antárticos (SC-ADM), el AMD es el directorio central de un sistema que contiene la descripción de los conjuntos de datos recogidos por los National Antarctic Data Centers (NADCs) que, a su vez, son repositorios financiados a nivel nacional (ANTARCTIC MASTER DIRECTORY, 2011a; SC-ADM, 2011a).

El repositorio AMD está alojado en el Global Change Master Directory (GCMD), a fin de reducir la duplicación de datos en el sistema. Ambos, AMD como los NADCs, integran la Antarctic Data Management System (ADMS) (SC-ADM, 2011b)[12].

[12] Más información disponible en < http://gcmd.gsfc.nasa.gov/KeywordSearch/amd/nadc_portals.html >.

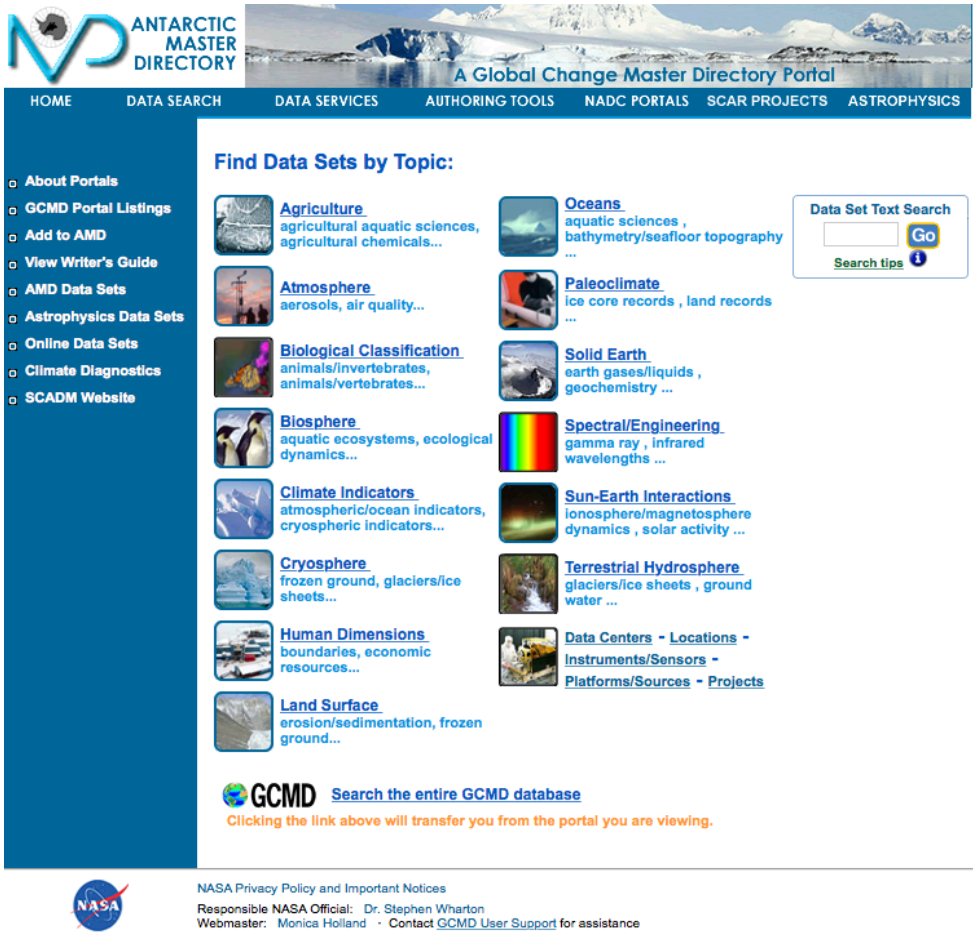

Figura 15: Datos generados por las encuestas nacionales en la Antártida
Fuente: Antarctic Master Directory (2014)

Los NADCs tienen repositorios donde se almacena los datos generados por las encuestas nacionales en la Antártida (SC-ADM, 2011a). Por lo tanto, los datos asignados por el investigador en NADC compondrán la red de información de AMD, que tendrá la información de todas las NADCs (SC-ADM, 2011b). Los NADCs que componen el AMD proceden de los programas de 21 países: Australia, Bélgica, Canadá, Argentina, Chile, España, China, Estonia, Finlandia, Francia, Uruguay, Italia, Japón, Corea del Sur, Malasia, Holanda, Nueva Zelanda, Reino Unido, Suiza, Ucrania y Estados Unidos (ANTARCTIC MASTER DIRECTORY, 2011b).

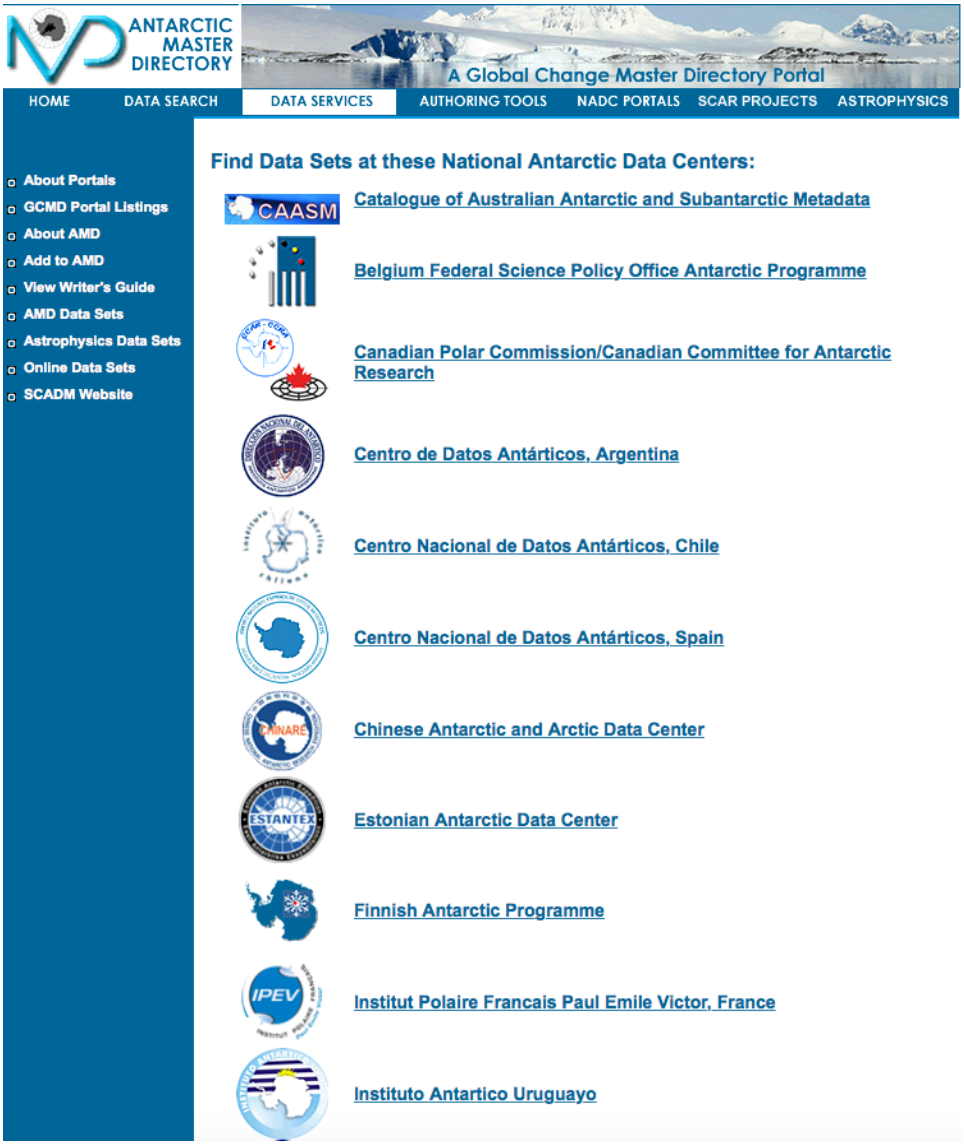

Figura 16: Los NADCs de los 21 países que componen el AMD
Fuente: Antarctic Master Directory (2014)

La gestión de datos del AMD está a cargo del comité SCAR SC-ADM, que se reúne anualmente para discutir la gestión de los datos en la investigación relacionada con la Antártida y estableciendo taxonomías de acuerdo a los principales centros de investigación en el estudio polar (SC-ADM, 2011c).

3.3 Áreas temáticas

Las áreas temáticas que conforman los estudios oceanográficos presentan actividades transversales de importancia crucial y atraviesan una amplia gama de ciencias del medio ambiente. Abarcan diversas disciplinas, incluyendo la

dinámica de los océanos (corrientes marinas, las olas y las mareas), la geología del fondo marino (forma, composición y formación del fondo), la composición química de los cuerpos de agua, los recursos minerales marinos; biodiversidad marina, los organismos y la ecología marina, por lo que se generan datos con la misma diversidad en cuanto a cobertura espacial, temporal, resolución y estructura de sus variables.

Los datos oceanográficos pueden ser de diferente naturaleza, y esto varía con los parámetros ambientales asociados con el estudio marino durante al momento de la recogida de datos. De acuerdo con el Repositorio Marino de Datos de Canarias[13], un análisis sistémico de datos y metadatos relevantes para una base de datos oceanográfica se puede explicar con la siguiente tabla:

DISCIPLINA	TIPO DE OBSERVACIÓN
Climatología marina	La temperatura y la humedad relativa; La presión de aire, sol, etc. La fuerza y la dirección del viento, el transporte, viento, el polvo, etc.
Física oceanográfica	La temperatura del agua, salinidad, turbidez, transparencia, etc.; Marine Dinámica: olas, corrientes, mareas, nivel del mar, etc.; Geofísica: geología, geomorfología fondos, cuevas, batimetría, granulometría, etc.
Oceanografía química	La química del agua (P, N, Fe y otros nutrientes); El oxígeno disuelto y clorofila del agua. La materia orgánica, el pH y la salinidad
Biodiversidad	Hábitats: tipología, Bionomics áreas degradadas bentónicos, etc. Especie: inventarios, avistamientos, varamientos, invasión, el desplazamiento (satélite, etc.). Conservación: Estado, el nivel de protección legal, etc. Los datos asociados con: fenología, biometría, ADN, condición, enfermedad, etc. Áreas de aves marinas. Los datos derivan: la producción orgánica. La concentración de las especies (medusas, etc.).
Arqueología	Las costas: basureros, playas levantadas, etc. Marina: naufragios, sitios arqueológicos, etc.
Medio ambiente	La contaminación del agua metales pesados, hidrocarburos, pesticidas, etc. contaminación microbiológica. La concentración de los residuos y la basura Los contaminantes del aire y aerosoles: NOx, SOx, los CFC, DMS, etc. Control de emisiones / emisiones de partículas en el aire y el ruido. Los parches de manchas de aceite. Las mareas rojas y floraciones de algas.

[13] Disponible en: < http://www.redmic.es >.

Datos políticos y administrativos	Límites administrativos: la zona económica exclusiva, aguas continentales, etc. Áreas Marinas Protegidas: ZEC, reservas marinas, parques, etc. Reservas pesqueras. Zonas portuarias: las zonas I y II de atraque, hay un área de acoplamiento, etc. Entidades (local): Port Authority, Police, rescue Marítima, La Cruz Roja, los pescadores, clubes de buceo, centros de investigación, etc. Contactos: usuarios, administradores, gerentes de proyecto, especialistas, etc. Documentación: proyectos, campañas, publicaciones, notas de expertos, etc.
El uso de los recursos	La pesca. Cultivos Marinos. Extracciones Otros: colección de especies, la extracción de arena, aceite, etc.
Navegación y incidentes	Las rutas marítimas. Campañas e investigación oceanográfica. Los transectos de estudio y observación
Imágenes	Imagen de satélite multibanda. Foto: Costa de ortofotografía, incidentes, especies, fondos, etc. Vídeo: trayectos submarinos, incidencias, etc.

Tabla 8: Tipos de observaciones realizadas en el ámbito marino según disciplina o ámbito de interés
Fuente: Repositorio Marino de Datos de Canarias

3.4 Formato de los datos oceanográficos

En la actualidad existen muchos formatos de datos y se van creando más a medida que se necesitan. Los mismos datos pueden aparecer en diferentes formas, con contenidos variados. De acuerdo con el COI (2007, p. 24) “no existe una estructura ´universal´ de datos, aunque hay indicios de una lenta convergencia hacia un pequeño número de estructuras de datos”. La falta de un conjunto universal de datos imposibilita el análisis diversificado de disciplinas para investigar un mismo tema.

El COI (2007) apunta la necesidad de cooperación entre los programas internacionales de gestión de datos para estimular una convergencia más rápida de las estructuras de datos. En especial es necesario establecer un conjunto más reducido de formatos de datos apropriados a las necesidades de almacenamiento por los repositorios de datos oceanográficos. Actualmente las bases de datos internacionales presentan numerosos formatos para el intercambio de datos, aunque algunos sean más comunes en las principales agencias de recolecta y divulgación.

A continuación presentamos los principales formatos de los datos seguido de una descripción de los más utilizados en investigaciones marinas:

Código	Nombre
ARCC	Coverage of Arc-Info
ARCE	ARC/INFO Export format
ARCG	ARC/INFO Generate format
ASCII	Formatted for text attributes
BIL	Imagery, band interleaved by line
BIP	Imagery, band interleaved by pixel
BMP	Windows or OS/2 Bitmap
BSQ	Imagery, band interleaved sequential
BUFR	Binary Universal Form for the Representation of meteorological data
CDF	Common Data Format
CFF	Cartographic Feature File (U.S. Forest Service)
COORD	User-Created Coordinate File
DBF	dBase File
DEM	Digital Elevation Model format (U.S. Geological Survey)
DFAD	Digital Feature Analysis Data (National Imagery and Mapping Agency)
DGN	Microstation Format (Intergraph Corporation)
DIGES	Digital Geographic Information Exchange Standard
DLG	Digital Line Graph (U.S. Geological Survey)
DTED	Digital Terrain Elevation Data (MIL-D-89020)
DWG	AutoCAD Drawing format
DX90	Data Exchange (90)
DXF	AutoCAD Drawing Exchange Format
ECW	ERMapper Compress Wavelets
ERDAS	ERDAS image files (ERDAS Corporation)
GRASS	Geographic Resources Analysis Support System
GRIB	GRIdded Binary or General Regularly-distributed Information in Binary form
GRID	Arc/Info Binary Format
HDF	Hierarchical Data Format
HTML	HyperText Markup Language

Código	Nombre
IGDS	Interactive Graphic Design System format (Intergraph Corporation)
IGES	Initial Graphics Exchange Standard
IMG	ERDAS Imagine format
JPEG	Joint Photographic Group Format
LAN	Earth Resources Data Analysis System
MDB	Microsoft Data Base
MIF	MIF
MOSS	Multiple Overlay Statistical System export file
NETCDF	Network Common Data Format
NITF	National Imagery Transfer Format
RPF	Raster Product Format (National Imagery and Mapping Agency)
RST	RST
RVC	Raster Vector Converted format (MicroImages)
RFV	Raster Vector Format (MicroImages)
SDTS	Spatial Data Transfer Standard (Federal Information Processing Standard 173)
SHP	ArcView ShapeFile
SIF	Standard Interchange Format (DOD Project 2851)
SLF	Standard Linear Format (National Imagery and Mapping Agency)
TAB	MapInfo Tabular Format
TGRLN	Topologically Integrated Geographic Encoding and Referencing (TIGER) Line format (Bureau of the Census)
TIFF	Tagged Image File Format
VPF	Vector Product Format (National Imagery and Mapping Agency)

Tabla 9: Formatos de datos oceanográficos
Fuente: Hernández-Jaimes (2008)

Los formatos más utilizados en investigaciones marinas, e indicados en la tabla 9, son los siguientes:

3.4.1 Formato de datos Jerárquicos HDF

HDF se desarrolló originalmente como un formato robusto, estándar para datos reticulados que varían en escalas de la superficie oceánica. Sigue siendo uno de los principales formatos para la distribución de los datos del Sistema de Observación de la Tierra (Earth Observing System - EOS) de la NASA, Estados Unidos.

El HDF es un formato estructurado de multiples objetos diseñado en el National Center for Supercomputer Aplications (NCSA) para facilitar la transferencia de datos entre maquinas diferentes.

3.4.2 Formulario de datos comunes de red (NetCDF)

El NetCDF es un formato abstracto para matrices multidimensionales diseñado en el Unidata Program Center, ubicado en Builder, Colorado. Este formato permite la representación de datos escalares, vectoriales así como de mallas irregulares, sin embargo solo los datos escalares, sobre mallados regulares, son importados de forma directa por el NetCDF. Para importar los otros tipos es necesario especificar algunos atributos extra en los datos.

Para importar datos escalares sobre malla regular se coloca en el módulo de importación el nombre del archivo y se especifica el formato "netCDF".

El NetCDF fue desarrollado principalmente para datos de la matriz (i.e. grids), pero se ha extendido a los datos de las mediciones, como se utiliza BUFR. Es ampliamente utilizado en la comunidad del clima, del tiempo y la marina, y hay indicios de que va a jugar un papel importante en los sistemas de observación del océano mundial emergentes. Recientemente el NetCDF 4.0 fue lanzado, incorporando HDF5, en representación de la primera unión de grandes formatos. NetCDF tiene un formato analógico ASCII, CDL, que puede ser fácilmente "compilado" a NetCDF.

El NetCDF está siendo utilizado de forma rutinaria en algunos programas globales de teledetección, por ejemplo, el Grupo de Alta Resolución temperatura superficial del mar (Group for High Resolution Sea Surface Temperature - GHRSST). Los archivos de cuadrícula NetCDF representan pocas dificultades para gestionarlos y son utilizados por una amplia variedad

de programas de visualización y análisis. Además, su uso favorece enormemente la compatibilidad entre los productos de datos y aplicaciones.

3.4.3 Formatos autodescriptivos

Los formatos autodescriptivos contienen de forma explícita los parámetros del dispositivo y la codificación en algún punto del fichero.

Estos formatos son los formatos operativos en uso hoy en día por la comunidad meteorológica mundial (GRIB, BUFR), la comunidad satélite (HDF) y los sistemas de observación de los océanos (NetCDF). Tres de ellos son adecuados para datos malla o raster (GRIB, HDF, NetCDF) y dos de ellos son adecuados para los informes de datos (BUFR, NetCDF). Contienen extensos metadatos internos, de ahí el nombre del grupo, proporcionando sistemas de usuario con toda la información necesaria tanto para la búsqueda de datos y el uso práctico. Los avances recientes que indican una fusión de estas tecnologías se indican a continuación.

WMO llama a las claves determinadas por tablas BUFR y GRIB, ya que requieren el uso de muchas tablas de códigos estándar (ver los códigos WMO de referencia más abajo). La comunidad meteorológica mundial ha llevado al desarrollo de estándares de datos, tales como tablas de códigos, y actualmente la comunidad oceanográfica ha comenzado a mirar hacia estos principios de la muestra.

3.4.4 CDF

El CDF es un formato abstracto para matrices multidimensionales auto-descriptivas un poco menos extenso que el formato nativo dx. Fue diseñado en el NASA/Goddar Space Flight Center.

Para importar este formato se debe colocar en el campo "name", del módulo de importación, el nombre del CDF (no el nombre del archivo, ya que el formato CDF permite el almacenamiento en multiples archivos).

3.5 Registro y calidad de los datos

Para la conversión de los conjuntos de datos a los formatos de difusión más comunes como TXT y CVS, se deben tener en cuenta los formatos en los que deben publicarse tanto la referencia temporal como la referencia espacial de los datos.

Para los campos de fecha y hora, el informe del foro sobre estándares, la International Oceanographic Data and Information Exchange (IODE, 2008), recomienda el uso de la norma ISO 86601:2004. Para los campos de latitud, longitud y altitud, recomienda la norma ISO 6709.

Cabe notar que los formatos TXT y CVS no son los únicos utilizados para el intercambio de datos marinos. Ortiz-Martinez, Mogollón Diaz y Rico-Lugo (2008) en la conferencia internacional sobre datos y sistemas de información marinos (IMDIS2008) muestran la experiencia de implementar en formato NetCDF (Network Common Data Form), un formato binario de gran acogida para la generación de archivos de datos físicos ya que permite incluir gran cantidad de meta-información en ellos, como por ejemplo, información del control de calidad, de valores máximos y mínimos esperados, instrumento de medición, etc, sin incrementar considerablemente el tamaño de los archivos.

Los datos redactados en tablas, merecen un especial cuidado con la precisión de su calidad; datos equivocados resultan una pérdida de dinero, tiempo y recursos.

Se pueden producir errores de tipográficos (digitación) y también en calidad de datos como la duplicación de registros, registros cercanos, que no son sintácticamente exactos pero que representan la misma entidad en el mundo real y donde un único identificador no está disponible.

La repetición de registros cercanos sucede cuando más de un registro corresponde a la misma persona, en un catálogo de clientes por ejemplo. La búsqueda de soluciones para este problema suscitó una tarea muy importante en el tema de calidad de datos siendo foco de muchas investigaciones en los últimos años. De acuerdo con Almeida (2001), la calidad de los datos se manifiesta en la descripción de un dato completo, consistente, exacto y

preciso; garantizando que los datos cumplan las necesidades de la investigación donde son utilizados.

Una solución ubicada en la literatura para corregir el problema de duplicación es la aproximación de registros, este método consiste en la comparación de registros en uno (reduplicación) o más conjuntos de datos con el esfuerzo de determinar qué pares de registros representan la misma entidad del mundo real (Elfeky & Verykios, 2002). El proceso de aproximación de registros solamente detecta los registros que se refieren a la misma entidad (posibles duplicados) no eliminando los mismos.

La evaluación de la calidad de los datos es el acto de juzgar en qué medida se puede confiar en que los valores observados representan lo que se está midiendo. La calidad de la investigación depende de datos de buena calidad y éstos a su vez dependen de métodos de control de buena calidad. El control de la calidad de los datos tiene el siguiente objetivo:

“Asegurar la coherencia de los datos dentro de un conjunto de datos y dentro de una colección de conjuntos de datos, y asegurar que la calidad y los errores de los datos son aparentes para el usuario que dispone de información suficiente para evaluar su idoneidad para una tarea determinada.” (COI, 2007)

La estrategia de gestión de datos e Información de la COI recomienda las mejores prácticas de control de calidad documentado (comprendida una serie estándar de pruebas de control de calidad automáticas), un control de calidad científico (aprobado por expertos idóneos) y un sistema único de rótulos de calidad, disponible y de fácil acceso.

Ortiz, Rubio (2007) señalan que la actividad de asignar banderas a los datos brinda al usuario información contundente y de primera mano sobre calidad del dato al que accede. Hoy por hoy se han adoptado gran variedad de convenciones de banderas de calidad, que certifican la calidad de los datos, las cuales generalmente se asocian al tipo de dato e instrumento de medición. Es importante tener en cuenta que los conjuntos de datos deben cumplir con los niveles de calidad mínimos exigidos por la comunidad científica y por esta razón es responsabilidad del proveedor de datos aplicar pruebas de calidad

rigurosas tanto a las variables temporales y espaciales del conjunto de datos, como a los datos oceanográficos y de meteorología marina, siendo ello un compromiso importante para el intercambio. Al concluir las pruebas de la calidad del dato deben ser asignadas banderas para informar al usuario acerca de la calidad del dato.

Las convenciones más reconocidas para calificar datos marinos se aprecian en la tabla 10 y dependiendo del recurso a describir se puede usar una u otra convención ya que no existe a la fecha un estándar global.

BANDERAS	CÓDIGOS DE CALIDAD	TIPO DE DATOS
1-9	WOCE HP (World Ocean Circulation Experiment)	CTD y muestras de botellas
0-9	IGOSS (Integrated Global Ocean Services System) y Mapping WOCE HP Codes to IGOSS	Datos IGOSS
0-9	World Ocean Database 2001 Quality Codes	Estaciones, observaciones y datos biológicos
0,1,4,8	Ocean Data View Quality Codes	- -
0,9	ARGO Quality Control Codes	Datos de boyas Argo
0,5	Data Quality Flags used in the UOT (Marine Environmental Data Services) and GTSPP (short version)	- -
1,7	IRD (Institut de Recherche pour le développement)	- -
0-9,A	Seadatanet	- -

Tabla 10: Banderas de calidad
Fuente: ORTIZ; RUBIO (2007)

3.6 Metadatos

Es necesario garantizar que los datos sean registrados, mantenidos y preservados de manera adecuada. Uno de los requisitos iniciales es que los conjuntos de datos estén acompañados de informaciones que describan cómo se han obtenido (tiempo o espacio, métodos y instrumentos de recogida), cuál es su ámbito, autoría, propiedad y condiciones de reutilización, control de calidad con el horizonte de un control por pares similar al que funciona en el caso del arbitraje de los artículos científicos, etc (Schaap &

Glaves, 2014). A este conjunto de descriptores se les denomina metadatos. Así, juntamente con la interoperabilidad tecnológica, la existencia de metadatos adecuados y normalizados es un requisito esencial para el acceso y reutilización de datos científicos.

Para el desarrollo de una base de datos en investigaciones oceanográficas es necesario enfatizar la importancia de los metadatos para una eficiente interoperabilidad, la cual permite el almacenamiento y la amplia divulgación para acceso a los datos estructurados. Estudios locales en bases de datos pueden formar parte de una amplia red de información, muchas veces internacional, lo cual aumenta el rango de la investigación, fomenta la publicación y expande la visibilidad de la producción científica en el área de interés. Por otro lado, reduce la posibilidad de replicación de los estudios de forma improductiva, lo que es equivalente a una apreciación adecuada de los financiamientos en ciencias.

Los metadatos oceanográficos son informaciones extraídas de investigaciones que representan una información documental que se desprenden del análisis de sus características básicas e intrínsecas. Se ensamblan en estructuras, dispuestas de acuerdo con el alcance de un área específica, la normalización y la descripción de los registros y la creación de patrones. Estas estructuras se extienden de cada uno de metadatos, por lo general bajo la responsabilidad de un centro de investigación o institución, adaptando su uso e intercambio de información entre los núcleos utilizando sistemas de metadatos compatibles. Así los metadatos presentan la información detallada de un recurso, generalmente expresada en textos y palabras-clave. Dicha información puede ser más compleja como un consenso de opiniones de variadas personas a respecto de un mismo recurso.

No hay un estándar de metadatos único y universal. Cada comunidad tiene sus propios requisitos para que describen los recursos y las diferentes normas (y los perfiles de estas normas) se han desarrollado para satisfacer las necesidades de la comunidad. Por ejemplo, el estándar Dublin Core se desarrolló para definir un conjunto de elementos que podrían ser utilizados por los autores para describir los documentos basados en la web. La ISO

19115[14] define un esquema para describir la información y servicios geográficos y proporciona información específica sobre los aspectos espaciales y temporales de los datos geográficos. Es importante recordar que los diferentes esquemas de metadatos se pueden utilizar para describir el mismo recurso y servir a un número de grupos de usuarios.

La Norma Internacional ISO 19115 es un estándar desarrollado por la International Organization for Standardization (ISO). Es un componente de la serie de normas ISO 191xx standards de metadatos geográficos, que define cómo describir información geográfica y servicios asociados, incluyendo contenidos, compras espacio-temporales, calidad de datos, acceso y derechos de uso. Los metadatos de la ISO 19115 distingue entre unos 20 elementos de metadatos básicos y define una lista completa de cerca de 400 elementos, con la mayor parte de ellas se enumeran como "opcional". Además, posibilita el desarrollo de comunidades individuales con un "perfil de la comunidad" de la Norma Internacional. Un grupo selecto de elementos de metadatos se puede establecer como obligatorio para una comunidad de usuarios. Una comunidad también puede establecer elementos de metadatos adicionales que no están en la Norma Internacional. Un perfil de la comunidad debe establecer tamaños de campo y dominios para todos los elementos de metadatos.

Las reglas para la creación de un perfil de la comunidad se describen en la norma ISO 19106 Norma Internacional de Información Geográfica (ISO 19106 International Standard Geographic Information). Esta norma define el modelo conceptual requerido para describir la información y servicios geográficos. El objetivo de esta Norma Internacional es proporcionar un procedimiento claro para la descripción de los conjuntos de datos geográficos digitales para que los usuarios sean capaces de determinar si los datos de una explotación serán de utilidad para ellos y cómo acceder a los datos. Mediante el establecimiento de un conjunto común de terminología de metadatos, definiciones y procedimientos de extensión, este estándar promove el uso adecuado y la recuperación de los datos geográficos. Los beneficios

[14] Explicaremos la norma ISO 19115 con más detalles en el capítulo 6.3.5

adicionales de esta norma son para facilitar la organización y gestión de los datos geográficos y para proveer información acerca de la base de datos de una organización para los demás. Ejemplos de perfiles de la comunidad de la norma ISO 19115 incluyen los perfiles como Marine Community Profile y el SeaDataNet Common Data Index.

Las descripciones de conjuntos de datos marinos son muy amplias y normalmente siguen una estandarización ISO. Los métodos utilizados para describir estos conjunto de datos a menudo han evolucionado de diferentes maneras y pueden ser incompatibles. Algunas descripciones de conjuntos de datos han sido desarrolladas a nivel nacional o regional para convertirse en un estándar, pero pueden no ser interoperable con otros países o regiones y no ofrece una manera fácil de acceder a descubrir los datos a nivel global.

A nivel de conformidad es importante asegurarse de que todos puedan descubrir, comprender y compartir los datos mediante la búsqueda y comparación de datos comunes con respecto a los datos. Esta normalización del vocabulario hace que el intercambio de información más fiable y universal. Los siguientes estándares de metadatos son algunos de los más importantes que se utilizan para describir los datos marinos.

La interoperabilidad permite que el contenido o los sistemas puedan trabajar juntos a través del uso de estándares que sean especificaciones acordadas. Es uno de los principios más importantes en la implementación de metadatos, ya que facilita el intercambio y puesta en común de los datos y permite la búsqueda de varios dominios, mientras que el intercambio de metadatos se ve facilitada por los cruces de información.

La elección adecuada del estándar de metadatos permite una descripción estandarizada eficiente, proporcionando la recuperación, la interoperabilidad entre los sistemas e intercambiar información de forma segura. Para seleccionar el mejor formato, debe investigar las diferencias entre los patrones, análisis de niveles de especificidad, la estructura y los objetivos de su desarrollo. De este modo, se asegura el nivel de calidad de la recuperación de la información por el repositorio, lo que garantiza la existencia de campos

específicos para incluir los metadatos de su área de cobertura.

De acuerdo con Rodrigues (2002), hay tres tipos de estándares de formato y los metadatos, como se puede ver en la siguiente tabla:

	Formato uno	Formato dos	Formato tres
Registro de funciones	Formato simple Estándar propietario Todo el texto indexados	Formatos estructurados Estándares emergentes Estructura de campos	Formatos altamente estructurados Las normas internacionales Estructura con etiquetas (tags)
Formatos de archivos	Google, Yahoo, etc.	Dublin Core, Planilha IAFA, RCF 1807, SOIF, LDIF	MARC, TEI, CIMI, EAD, ICPSR

Tabla 11: Tipología de los formatos de metadatos.
Fuente: Rodríguez (2002)

Los tres niveles de las bandas descritas en la tabla anterior tienen diferentes tipos de estándares de metadatos, que van desde una estructura simple, intermedia, a una descripción más compleja. Estas características de los formatos de metadatos tienen particularidades muy diferentes, como se muestra a continuación:

Tipo de formato	Principales características	Ejemplo
Metadatos simples	Están organizados por metadatos no estructurada, extraída de forma automática, por lo general con la semántica reducidas	*MetaTag(s)* y metadatos utilizados en la transferencia de datos utilizando el protocolo *Hipertext Transfer Protocol* (HTTP)
Estructurado	Junto con los metadatos marco basado en estándares con una descripción mínima del recurso para la identificación, localización y recuperación de datos; La descripción se lleva a cabo generalmente en el campo y esta categoría requiere la ayuda de expertos	Estándar de Metadatos *Dublin Core*

Altamente estructurados	Están compuestos de complejo marco de descripción de metadatos y presente en el nivel más detallado; Basado en estándares y códigos especializados de un dominio particular proporcionar a la descripción de una fuente de información o individuo que pertenece a una colección y facilitar la localización, recuperación, intercambio de recursos de información	Formato MARC estándar 21

Tabla 12: Características de los formatos de metadatos
Fuente: Elaboración propia con base en el estudio de Alves (2005)

Las características de los tipos de metadatos presentadas en último cuadro muestra rasgos distintivos de su uso en entornos digitales, pero es esencial para comprender las razones de la necesidad de la normalización son cada vez más importantes.

Los metadatos no solo auxilian el descubrimiento de los recursos sino también ayudan al usuario a evaluar su utilización para determinado propósito. Existen varios estándares de metadatos utilizados para organizar los documentos oceanográficos, dependiendo de la finalidad de estos. A continuación se presentan algunos ejemplos más representativos en el medio oceanográfico, empezando por el Dublin Core (DC). El DC tiene ventaja sobre otros formatos, debido a la responsabilidad de su desarrollo por el W3C, y se recomienda como un estándar de uso de Internet, ya que cumple con las necesidades específicas de la web. Una de las principales características de la DC es para describir recursos electrónicos de una manera simplificada, proporcionando una base para la interoperabilidad semántica con otros formatos ampliamente utilizados y seguir las normas internacionales para la descripción de la información electrónica.

3.6.1 Dublin Core (DC)[15]

[15] http://dublincore.org

El estándar Dublin Core, presentado por Dublin Core Metadata Initiative (DCMI)[16], el Resource Description Framework (RDF) [17] y por el World Wide Web Consortium (W3C)[18], fue creado para resolver la descripción del problema y la recuperación de los recursos electrónicos en la web. Para expandir sus operaciones, el DCMI ha desarrollado la semántica de la DC en términos generales con las cuestiones de sintaxis definidos. Por otro lado, las reglas sintácticas RDF desarrollados en los que el DC estándar puede ser incorporado, proporcionando una integración útil de ambos. Por otra parte, mientras que la componente de corriente continua no es un RDF directo, uno de los primeros regímenes evaluados y utilizado es RDF. Típicamente, los descriptores de DC se construyen en el objeto digital (HTML, XML, etc.), pero pueden ser grabados en repositorios separados.

Mediante la adopción de la norma DC, la infraestructura recomendada por la W3C prevé el intercambio de metadatos con otras aplicaciones web, porque no es un lenguaje consolidado por el W3C para expresar metadatos DC, el lenguaje RDF/XML. La sintaxis de RDF se basa en el lenguaje de marcado XML[19], establecida por la W3C y el gobierno brasileño, como se indica en la versión 0 de la arquitectura de los Padrões de Interoperabilidad del Gobierno Electrónico (e-PING)[20] (COMITÊ EXECUTIVO DE GOVERNO ELETRÔNICO, 2004).

---

16 http://dublincore.org

17 Significa Resource Description Framework, es un modelo de datos estructurados en los gráficos y cuenta con varios formatos de serialización como RDF / XML, Notación 3, Turtle. Formatos de base tienen sus datos RDF descritos en vocabularios disponibles en la Web A pesar de la gran calidad de los datos disponibles en RDF, la construcción de vocabulario para su uso no es trivial. En una escala de niveles de calidad/complejidad de los datos abiertos, RDF es el último nivel, que constituyen la Web semántica.

18 http://www.w3.org

19 Soportes de Extensible Markup Language. Se trata de un conjunto de reglas de codificación para documentos en una estructura jerárquica y formato legible por la máquina. Es basado en texto y tiene como objetivos principales la simplicidad, extensibilidad y facilidad de uso. XML se utiliza ampliamente como un formato de intercambio de datos de los clásicos servicios web SOAP. Cuenta con una amplia gama de herramientas asociadas, como el XSLT predeterminado para la transformación de XML a otra estructura u otro formato.

20 Los Estándares de Interoperabilidad de Gobierno Electrónico (e-PING) permiten un flujo continuo de información entre el gobierno y la sociedad, ayudando a proporcionar mejores servicios a los ciudadanos. Esta arquitectura permite a los mismos sistemas de información con diferentes arquitecturas y desarrolladas en diferentes momentos para generar e intercambiar información en tiempo real. El documento está disponible en www.eping.e.gov.br

El estándar de metadatos Dublin Core es un conjunto de elementos para describir una amplia gama de recursos en red, desarrollado por la editorial, la biblioteca y las comunidades académicas, y se centra en las necesidades bibliográficas. El Conjunto de Elementos de Metadatos Dublin Core, la versión 1.1 incluye quince elementos de metadatos para su uso en la descripción de recursos. Estos elementos "básicos" son amplios y genéricos, y se pueden utilizar para describir una amplia gama de recursos.

3.6.2 Estándar de metadatos sobre biodiversidad

En el proyecto de estandarización de información biológica en las Américas que se incorpora al Sistema de Información sobre Biodiversidad de Colombia (SIB) se desarrolló un estándar de metadatos sobre biodiversidad que facilita la transcripción de datos taxonómicos provenientes de registros biológicos (Humboldt, 2003) y por lo tanto marca campos y vocabularios controlados especializados sobre información taxonómica.

3.6.3 FGDC (Federal Geographic Data Comittee)

Estándar de metadatos utilizado para la descripción de datos geoespaciales. Este formato ha sido requerido de todas las agencias de Estados Unidos; debido a su tamaño (unos 300 campos) no es del todo popular.

3.6.4 IAFA/WHOIS++(Internet Anonymous Ftp Archive with Whois++ protocol)

Estándar de metadatos utilizado para descripción del contenido y servicios disponibles en archivos FTP (File Transfer Protocol).

3.6.5 Marine Community Metadata Profile

Desarrolado por el Centro Nacional de Datos Oceanográficos de Australia (AOCD), el perfil de metadatos marinos ha alcanzado una completa compatibilidad con el estándar de metadatos geográfico ISO19115:2003, se han definido elementos obligatorios, elementos suplementarios, listas de

código y vocabulários controlados para asistir la descripción de los recursos marinos.

3.6.6 SAIF (Spatial Archive and Interchange Format)

Estándar de metadatos para la caracterización de datos espaciales e espaciotemporales.

3.6.7 VMO Core (World Meteorological Organization)

Es el perfil del estándar de metadatos geográfico ISO19115:2003 utilizado para la descripción de datos meteorológicos.

Muy seguramente adoptar ISO19115 o un perfil basado en él asegurará la interoperabilidad entre los sistemas de información geográficos marinos tal como lo señala COI, 2008, no obstante, todavía queda una gran cantidad de servicios e información más heterogéneos susceptibles se ser ofrecidos a través de una Infraestructura de Datos Espaciales (IDE); para el caso que nos atañe, por ejemplo la producción científica que reposa en la biblioteca de esta institución.

Actualmente Dublin Core se ha convertido en una parte importante en la caracterización de recursos para internet debido a su simplicidad, es por ello que Ortiz-Martine y Mogollón Diaz (2008) reportan en su revisión el perfil de aplicación de metadatos SDIGER[21] - Dublin Core para minería de datos geográficos cuyo punto de partida es el “Dublin Core Spatial Application Profile” definido por el CEN/ISSS Workshop on Metadata for Multimedia Information - Dublin Profile” definido por el CEN/ISSS Workshop on Metadata for Multimedia Information - Dublin Core (WS/MMI-DC), señalando así la posibilidad de desarrollar un perfil de metadatos marinos ISO-19115 - Dublín Core, el cual es parte de los objetivos de la fase II del proyecto de implementación de la central de información Marina Colombiana (CENIMARC) en los centros de investigación de la Dirección General Marítima.

---

21 SDIGER es un proyecto piloto para la implementación de la Infraestructura Europea de Datos Espaciales (Infraestructura for Spatial Information in Europe, INSPIRE), financiado por la Comisión Europea a través de Eurostat

3.6.8 El Marine Community Profile (MCP)

El Marine Community Profile (MCP) de la norma ISO 19115 se ha desarrollado de acuerdo con las reglas establecidas por la norma internacional con el auspicio del Fondo para el Australian Ocean Data Centre Joint Facility (AODCJF). El MCP incluye en todas las normas ISO 19115 elementos de metadatos núcleo y elementos no esenciales seleccionados. El MCP también ha definido los elementos de metadatos complementarios y listas de códigos para satisfacer las necesidades de la comunidad marina para apoyar la documentación y el descubrimiento de los recursos marinos. La documentación para el MCP está disponible por medio de la AODCJF.

3.6.9 El Common Data Index (CDI)

El Common Data Index (CDI) proporciona una visión detallada de los conjuntos de datos disponibles y allana el camino para dirigir el acceso de datos en línea o las solicitudes en línea directa para el acceso de entrega de datos. El principio del CDI es que cada centro de datos participante produzca a intervalos regulares. Las contribuciones de los asociados y sus actualizaciones periódicas se recogen en el centro de una base de metadatos CDI central, que está equipado con una interfaz de usuario CDI, para servir a los usuarios. CDI ha adoptado XML y la norma ISO 19115 para apoyar el intercambio estándar y la interoperabilidad. CDI documentación está disponible desde SeaSearch.

3.6.10 Sample ISO 19115 Records[22]

Se trata de un registro de metadatos que describe datos de CTD. El registro contiene todas las estaciones CTD medidos durante un crucero (que se muestra en formato XML).

3.6.11 SDIGER - WFD

Es un perfil de aplicación de metadatos de ISO19115 para la descripción de los recursos exigidos por la Directiva Marco del Agua. De este modo, este

[22] Explicaremos al formato ISO 19115 con más detalles en el apartado 3.6.5

perfil se basa principalmente en la guías para metadatos incluidas en el documento “Guidance Document on Implementing the GIS Elements of the Water Framework Directive”. Adicionalmente han sido tenidas en cuenta otros estándares e iniciativas relacionadas con metadatos y aspectos medioambientales.

3.6.12 Directorio Interchange Format (DIF)

El Directory Interchange Format (DIF) es un estándar de metadatos utilizado por la NASA Global Change Master Directory (GCMD) para describir conjuntos de datos de ciencias de la Tierra. Cuenta con un total de 36 elementos, entre ellos 8 elementos obligatorios que son EntryID, Título de entrada, palabras clave, ISO Categoría Tema, Data Center, Resumen, Metadatos Nombre y Metadatos Versión. Algunos de los campos son campos de texto, otros requieren el uso de palabras clave controlados (a veces conocido como "validas"). El estándar DIF es compatible con las normas ISO 19115 y CSDGM.

3.6.13 Geonetwork

Para la documentación de recursos compatibles con el estándar de metadatos ISO-19115 se puede contar GeoNetwork, una herramienta libre y de código abierto basada en estándares para manejar recursos espaciales, desarrollada por la Organización de las Naciones Unidas para la Agricultura y la Alimentación (FAO-UN), el Programa Mundial de las Naciones Unidas para la Alimentación (WFP-UN) y el Programa de las Naciones Unidas para el Medio Ambiente (UNEP).

Geonetwork implementa tanto el componente del portal web como el catálogo de metadatos para bases de una Spatial Data Infrastructure (SDI) definida en la arquitectura de referencia OGC (Open Geospatial Consortium). Dentro de sus alcances se destaca la capacidad de realizar búsquedas distribuidas que proporcionan acceso a un volumen enorme de metadatos de diversos Clearinghouses[23] y también proporciona un mapa interactivo que permite

[23] Agencias asociadas con el intercambio, comercio y divulgación de recursos.

escoger diversas capas de información de los servidores distribuidos en el internet.

3.6.14 Resumen de reportes de cruceros (CSR)

La RSE, anteriormente conocido como ROSCOP (Report of Observations/ Samples Collected by Oceanographic Programmes - Informe de Observaciones / muestras recogidas por los Programas Oceanográfico), es un estándar internacional establecido diseñado para recopilar información acerca de los datos oceanográficos. ROSCOP fue diseñado a finales de 1960 por el COI para proporcionar un inventario para el seguimiento de los datos oceanográficos recogidos en buques de investigación. Fue revisado extensamente en 1990 y pasó a llamarse Resumen de reportes de cruceros (CSR-Cruise Summary Report). Ha sido ampliamente adoptado por muchos Estados miembros de la COI y otras organizaciones, como el Consejo Internacional para la Exploración del Mar (ICES - Council for the Exploration of the Seas)

3.6.15 Contenido Estándar para Metadatos Geoespaciales Digitales

El Content Standard for Digital Geospatial Metadata (CSDGM), a menudo referido como el estándar de metadatos FGDC, fue desarrollado y es mantenido por el Federal Geographic Data Committee de EE.UU. y es el estándar oficial de metadatos del país, incluyendo cerca de 300 campos obligatorios y opcionales.

3.6.16 Cruise Summary Report

El Cruise Summary Report (CSR), anteriormente conocido como ROSCOP (Report of Observations/Samples Collected by Oceanographic Programmes) muestra recogidas por los Programas Oceanográficos, es un estándar internacional establecido para recopilar información acerca de los datos oceanográficos. ROSCOP fue concebido a finales de 1960 por el COI para proporcionar un inventario de bajo nivel para el seguimiento de los datos oceanográficos recogidos en buques de investigación.

La forma ROSCOP fue revisado extensamente en 1990, y fue rebautizada como CSR (Cruise Summary Report), pero el nombre ROSCOP persiste para

muchos científicos marinos. La mayoría de las disciplinas marinas están representadas en ROSCOP, incluyendo física, química, y la oceanografía biológica, la pesca, la contaminación marina/contaminación, y la meteorología marina. La base de datos ROSCOP se mantiene por el ICES.

3.6.17 Muestra CSR Record

Este es un ejemplo de un registro de metadatos de CSR para el crucero de Bélgica en el Mar del Norte.

3.7 Ciclo de vida de los datos oceanográficos

El ciclo de vida de los datos oceanográficos incluye todas las actividades que afectan la vida útil de un conjunto de datos y pueden ser evidenciados en tres grupos, como se muestra a seguir:

Planificación y Producción	Incluye todas las actividades desde el momento en que una observación es capturada por un sistema o una colección de datos del proyecto de observación
Gestión de datos	Incluye todas las actividades relacionadas con el procesamiento, verificación, documentación, publicidad, distribución y conservación de los datos
USO	Incluye todas las actividades realizadas por parte de los investigadores (estas actividades a menudo quedan fuera del control directo de los administradores de datos)

El ciclo de vida de los datos oceanográficos es un proceso dinámico que no sigue una secuencia lineal. Es decir, los pasos en el ciclo de vida no son independientes, pues dependen de la influencia y las acciones tomadas en otros pasos. Por eso, la documentación inadecuada en la recogida de datos de campo pueden evitar su uso posterior. La generación de productos a partir de datos originales pueden producir nuevos datos derivados que también deben ser recogidos y gestionados, mientras que los comentarios de los usuarios respecto a los datos pueden cambiar o aumentar la documentación sobre ellos.

Por lo tanto, el ciclo de vida de los datos oceanográficos debe asegurar que los registros se basen en requisitos que posibiliten ser utilizados tanto para su propósito original a reutilizados para otros fines.

Cada fase del ciclo de vida de los datos oceanográficos se describe en las siguientes sub-secciones:

3.7.1 Planificación de requisitos: Presenta tareas para la evaluación de las necesidades y requerimientos de una investigación oceanográfica, la planificación de cómo satisfacer esos requisitos y la forma de gestionar los datos resultantes, el desarrollo de los sensores necesarios, implementación y la operación del sistema de observación.

La planificación incluye la preparación para la gestión de los datos y la directiva de procedimientos que requiere dicha planificación proporcionando una plantilla de preguntas. Esta planificación debe ser flexible y actualizada, pues los asuntos no considerados en el plan original o cambios en la tecnología pueden alterar la manera como se procesan, distribuyen y son archivados los datos. Por lo tanto, los administradores de programas, jefes de proyecto y personal técnico deben trabajar juntos y con los grupos de investigación para planificar la gestión de manera que maximicen la compatibilidad de datos oceanográficos y reduzcan los costos generales.

Las otras actividades de esta fase están en gran medida fuera del alcance de este marco y se centran en la gestión de los datos reales una vez que las observaciones son recogidas. Sin embargo, las actividades que se producen más tarde en el ciclo de vida de datos pueden influir en esta fase. Por ejemplo, un error de calibración descubierto durante el control de calidad puede llevar a cambios en el procedimiento de operación y el análisis de brechas puede revelar nuevos requisitos.

3.7.2 Gestión de datos

El procesamiento de un volumen de datos del medio marino en constante aumento plantea una serie de problemas tecnológicos que sólo han podido

ser resueltos con el uso de la tecnologia de la información. Las tecnologías de procesamiento de datos y los sistemas de detección y transmisión han avanzado enormemente. Con el objetivo de abordar de forma coordinada las actividades encaminadas a la preservación de los datos, la gestión de los datos requieren una serie de procesos que presentamos en la secuencia:

3.7.2.1 Recopilación de datos

Se refiere a los pasos iniciales de la recepción de datos en bruto de un sensor ambiental o una campaña de observación. La recolección también puede incluir la compra de bases de datos comerciales, la negociación de acuerdos de acceso a los datos de los sistemas extranjeros, la emisión de contratos para la recopilación de datos y la emisión de becas de investigación que pueden resultar en la creación de datos ambientales. Las estrategias para el diseño de los programas de recopilación de datos variarán según las políticas de cada país, cada un tendrá sus características, su propia importancia relativa y sus propias posibilidades de proporcionar datos. Además, puede ser necesario obtener datos de fuentes externas, como los datos relativos las investigaciones fuera del ámbito de la investigación prevista en el plan de gestión de los datos.

Es fundamental tener la información sobre la infraestructura para elaborar el marco de un programa de recopilación de datos. El primer paso es definir qué zonas se incluirán y preparar una descripción de las actividades de investigación que se efectuarán en ellas. Tal información sirve para proporcionar una clasificación y una descripción detallada que será fundamental para establecer un buen programa de recopilación de todos los datos relativos a la investigación. Muchos de estos datos también servirán para nuevas análisis y investigaciones.

3.7.2.2 Procesamiento de datos

El procesamiento de datos incluye todos los pasos necesarios para utilización de los datos en registros con posibilidades de generar métodos habituales de

consulta. Este tratamiento se realiza normalmente mediante sistemas especializados que tienen sus propios controles internos de gestión de datos. Los usuarios normalmente no tienen acceso directo al sistema de procesamiento. Sin embargo, el diseño de estos sistemas puede tener un gran impacto en los costes, la preservación y la calidad de los registros de datos y productos resultantes. La UNESCO (2007) recomienda aprovechar las inversiones pasadas o los recursos existentes para que los sistemas de procesamiento no sean construidos desde cero para cada plan de gestión de los datos oceanográficos.

3.7.2.3 Control de calidad

Los datos deben ser de calidad conocida, lo que significa que la documentación debe incluir el resultado de los procesos de normas y control de calidad para proporcionar la validación sobre los requisitos previstos en todo el ciclo de vida de los datos. El control de calidad puede incluir, por ejemplo, la calibración de datos de sensores en varios sistemas. Los resultados de estos controles deben incluirse en los metadatos como estimaciones de error o abanderamiento de los valores malos o sospechosas. Los datos en bruto que no han sido sometidos a control de calidad deben estar claramente documentados como de calidad desconocida.

3.7.2.4 Documentación

La Documentación de datos proporciona información sobre la extensión espacial y temporal, fuente, linaje, responsables, atributos descriptivos, la calidad, la precisión, la madurez, las limitaciones conocidas y la organización lógica de los datos. La documentación formal estructurada es posible por intermedio de los metadatos y son fundamentales para documentar y preservar los activos de datos oceanográficos. Los estándares de metadatos posibilitan la interoperabilidad de apoyo con catálogos, archivos y herramientas de análisis de datos para facilitar la búsqueda y el uso de datos. Los metadatos correctos y completos son esenciales para asegurar que los

datos sean utilizados de manera adecuada y que todos los análisis resultantes sean creíbles.

Los estándares de metadatos centrales para los datos oceanográficos en general son el modelo ISO 19115 (contenido) e ISO 19139 (Extensible Markup Language [XML]), mientras que el modelo para uso dependerá de lo que sea establecido por la directiva de procedimientos de documentación de datos prevista en la *planificación*. Algunos registros de metadatos utilizan los estándares Federal Geographic Data Committee (FGDC) y Standard for Digital Geospatial Metadata (CSDGM) y son convertidos a la norma ISO. La conversión de metadatos bien estructurados (por ejemplo, en FGDC XML) para ISO es relativamente sencillo, pero la documentación de forma libre no estándar es más problemático.

3.7.2.5 Catalogación

Se refiere a todos los mecanismos establecidos por los proveedores de datos para permitir a los usuarios encontrar los datos. Los datos oceanográficos deben ser fácilmente reconocibles porque la toma de decisiones dependenfundamentalmente de la capacidad de encontrar datos relevantes de múltiples agencias y disciplinas.

Los métodos de catalogación permiten el establecimiento de servicios de catálogo basados en estándares formales, portales web temáticos o del gobierno. La búsqueda general en la web es a menudo el primer paso para los usuarios potenciales, por lo que este servicio debe ser apoyado. Sin embargo, la búsqueda avanzada basada en la ubicación, el tiempo, la semántica u otros atributos de datos requieren catálogos de servicios formales.

La proliferación de portales como data.gov, geo.data.gov, ocean.data.gov, la NASA Global Change Master Directory (GCMD), el Group on Earth Observations (GEO) y otros proveedores de datos requiere múltiples registros en diferentes sitios. Este procedimiento conduce a esfuerzos redundantes y

una catalogación duplicada. Los proveedores de datos deben ser capaces de registrar su servicio en un único catálogo mientras otros catálogos y portales se conviertan automáticamente al tanto de los nuevos datos.

3.7.2.6 Difusión

Se refiere a la transmisión de datos de forma activa, la más típica, permitiendo a los usuarios acceder a los datos por encargo. Los datos oceanográficos deben ser fácilmente accesibles a los usuarios potenciales. Muchos usuarios prefieren el acceso directo a los datos en línea a través de los servicios de Internet que permiten las peticiones personalizadas en lugar de la descarga de archivos estáticos o el acceso diferido a través de solicitación de los servicios de datos que no están disponibles automaticamente en línea. Para la recopilación de datos de gran volumen que requieren almacenamiento no lineal, los administradores de datos deben considerar cuidadosamente las estrategias de *cloud hosting* basado en el seguimiento de uso para maximizar la probabilidad de los datos que son populares en línea. Los servicios en línea deben cumplir con las especificaciones de interoperabilidad para los datos geoespaciales, en particular los de OGC, ISO / TC211 y Unidata.

En algunos casos es necesario transmitir datos de forma activa a los usuarios operacionales, es decir, los usuarios que necesitan datos de manera no solamente frecuente, como también instantánea para avanzar en el desarrollo de sus investigaciones. Sin embargo, el establecimiento de nuevos conductos de datos que son de propiedad o duplicación debe ser evitado y los canales de distribución existentes deben ser compartidas siempre que posible. El hardware y el software utilizados para acceder a los datos deben ser utilizados de acuerdo con las necesidades de los usuarios. Tecnologías de código abierto financiados por el gobierno deben ser consideradas.

Los datos deben ser ofrecidos en formatos que sean conocidos por trabajar con una amplia gama de herramientas científicas o apoyo a las decisiones. Vocabularios comunes, la semántica y modelos de datos también deben ser empleados.

A menudo se difunden a los usuarios resultados de modelos numéricos, por ejemplo, obtenidos por satélite, sensores; como también datos observacionales, por ejemplo, por medio de fotos. Siempre que sea posible, los servicios y formatos compatibles con los datos, sean observacionales, numéricos o de otro tipo deben ser difundidos para facilitar la integración o la comparación de los datos y resultados de los modelos de varias fuentes.

3.7.2.7 Conservación y manejo

La preservación de datos asegura que los datos sean almacenados y protegidos de pérdidas. La custodia asegura que los datos sigan siendo accesibles (por ejemplo, mediante la migración a las nuevas tecnologías de almacenamiento) y se actualizan, anotando o reemplazando cuando hay cambios o correcciones. La administración también incluye el reprocesamiento cuando los errores o sesgos han sido descubiertos en el procesamiento inicial.

En el caso de los Centros Nacionales de Datos, tales como el NOAA, NGDC y el NODC - sus datos son operados por el National Environmental Satellite, Data, and Information Service (NESDIS), donde realizan la preservación de datos y la administración en nombre de toda esta agencia. Los productores de datos de la NOAA establecen un acuerdo de sometimiento con uno de estos centros de datos como se describe en el *Procedure for Scientific records appraisal and archive approval* (NOAA, 2016). Para asegurar que los datos producidos por los beneficiarios sean archivados, la Federal Funding Opportunity (FFOs) debe arreglar y presupuestar de antemano junto a un centro de datos de la NOAA el coste para el archivo de datos que se producido por los investigadores financiados.

Debido a que una observación no se puede repetir una vez que el momento ha pasado, todas las observaciones deberán archivarse. No sólo se deben preservar los datos en bruto, sino también la información de acompañamiento necesarias para la comprensión de las condiciones en el momento de la observación. En algunos casos, especialmente en el caso de las imágenes de

satélite de alta resolución, la falta del estricto cumplimiento de este principio daría lugar a costes adicionales sustanciales a las redes de telecomunicaciones y sistemas de almacenamiento de datos. La representación de los datos que debe ser preservado y manejado a largo plazo, debe ser negociada con el Centro de Datos e identificado en el plan de gestión de datos apropiado. También deben ser preservados productos derivados clave, o las versiones pertinentes de los programas informáticos necesarios para regenerar los productos que no se archivan.

El rescate de datos se refiere a la conservación de los datos que se encuentran en riesgo de pérdida. Estos datos incluyen información registrada en papel, película u otros medios obsoletos, o que carecen de metadatos esenciales, o almacenado sólo en el ordenador del científico. El rescate de datos es mucho más caro que asegurar la preservación de los conjuntos de datos actuales. En el caso de la de la NOAA, los conjuntos de datos en riesgo deben estar registrados en el *International Council for Science* (ICSU), por medio del *Committee on Data for Science and Technology* (CODATA).

Los datos que han sido enviados a un centro de datos de la NOAA también deben ser visibles y accesibles como se describe en las secciones anteriores. Idealmente, los mecanismos para la catalogación y difusión de los datos de archivo deben ser interoperables con los datos adquiridos en tiempo casi real.

3.7.2.8 Retención de registros

Los centros nacionales de datos normalmente cuentan con un programa de retención de registros que documenta el tiempo que los datos serán conservados. Los productores de datos también deben tener un programa de retención de registros que indica cuando los datos deben ser transferidos a un centro de datos para la conservación a largo plazo. Para eso la consolidación de los recursos de TI será cada vez más necesaria para transferir la custodia de los registros de datos desde los servidores locales a los centros nacionales de datos oceanográficos.

3.7.3 Uso

Se refiere a la capacidad para medir la frecuencia de los conjuntos de datos que se están utilizando, mientras la estimación bruta se puede hacer contando las solicitudes de datos o los volúmenes de transmisión de datos desde los servidores de Internet. Sin embargo, estas estadísticas no revelan si los datos que se obtuvieron se utilizaron en realidad, si lo utilizado era útil, o si el destinatario inicial redistribuye los datos a otros usuarios.

Es por ello que medios más sofisticados de evaluación de uso, preservando el anonimato de los usuarios son deseables. Los productores de datos deben asignar identificadores persistentes para cada conjunto de datos, e incluir el identificador en cada archivo de registro de conjuntos de datos.

3.7.3.1 Actividades de uso

La tercera fase del ciclo de vida de datos oceanográficos se refiere al uso de los datos. Estas actividades están normalmente fuera del alcance del administrador - una vez que un usuario haya obtenido una copia de los datos deseados, no es posible controlar el uso que el usuario pueda hacer de ellos. Sin embargo, la capacidad de obtener y utilizar los datos es ciertamente un subproducto de un proceso de gestión de datos del buen ciclo de vida, y la información acerca de los usuarios pueden influir o mejorar el proceso de gestión de datos. De acuerdo con las informaciones del NOAA (2016) el portal es el mayor usuario de sus propios datos, por lo que las mejoras en la gestión de datos podrían reducir el costo y la complejidad dentro de la agencia.

3.7.3.2 Análisis de las deficiencias

El *análisis* se define ampliamente para incluir actividades tales como una evaluación rápida para la utilidad de un conjunto de datos, o la inclusión de un conjunto de datos entre los factores que conducen a una decisión, o un

análisis científico real de datos en un contexto de investigación, o la minería de datos. Estas actividades son posibles si los datos han sido bien documentados y son de calidad conocida.

Los usuarios de los datos oceanográficos pueden crear productos derivados de valor agregado para constituir un nuevo conjunto de datos originando su propio proceso de gestión del ciclo de vida de los datos. Proyectos que crean rutinariamente nuevos productos establecen y siguen un plan de gestión de datos y aseguran que los productos que generan sean visibles, accesible y archivados. Nuevos productos deben vincularse de nuevo a los datos de origen originales a través de la documentación y la citación de los identificadores de conjuntos de datos correspondiente.

Los usuarios deben tener un mecanismo con facilidad de uso para proporcionar información con relación a las cuestiones sospechosas de calidad y otros aspectos de los conjuntos de datos. Además, es recomendable que haya la posibilidad de inclusión de nuevos metadatos por los investigadores con el fin de ayudar a los futuros usuarios. La inclusión de metadatos aumenta la capacidad para hacer referencia de forma inequívoca a un conjunto de datos y amplia la capacidad de identificar un conjunto de datos sin necesidad de modificar los metadatos originales. El análisis de los usuarios de datos o de los tomadores de decisiones que indican la necesidad de incluir datos adicionales para satisfacer las necesidades operacionales, por ejemplo, de la cobertura más frecuente, mejoraran la resolución espacial o espectral, o las observaciones de otras cantidades. Análisis de las lagunas también puede abordar la continuidad de las observaciones para satisfacer las necesidades operacionales o habilitar el análisis de tendencias a largo plazo. Tal determinación influye en la definición de requerimientos, que es el inicio de un nuevo ciclo de vida de datos.

3.8 Los modelos conceptuales de gestión de datos

Son los modelos orientados a la descripción de estructuras de datos y el flujo de datos entre provedores. Se usan fundamentalmente para definir la estructura de intercambio entre los centros de datos y están orientados a

representar los elementos que intervienen en la escogida de una arquitectura y sus relaciones.

3.8.1 Modelo centralizado

De acuerdo con la Intergovernmental Oceanographic Commission de la UNESCO (2016), la elección del modelo a adoptar se basará en una variedad de consideraciones. Por ejemplo, en la adopción de un modelo centralizado, los proveedores (investigadores, proyectos, campañas de investigación, plataformas de observación) no tienen la capacidad de gestión de datos. Todos los datos son enviados a un centro nacional de datos oceanográficos.

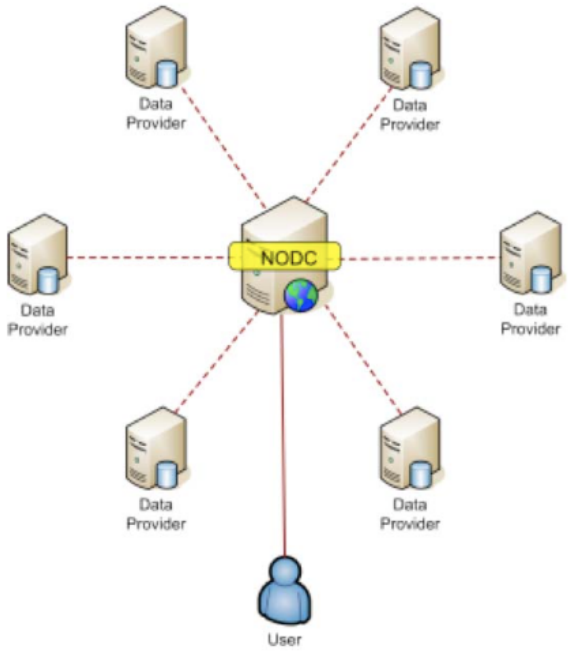

Figura 17: Modelo centralizado de centro de datos.
Fuente: Intergovernmental Oceanographic Commission (2016)

En el modelo centralizado, los datos se almacenan el almacenamiento puede atender a varios usuarios, pero la base de datos en sí residen por completo en una sola máquina o servidores centralizados localmente.

3.8.2 Modelo distribuido

En relación al modelo centralizado, el modelo distribuido de centro de datos tiene varias ventajas potenciales. Una ventaja es que los datos pueden ser manejados directamente en el centro de datos. Por ejemplo, si el centro de datos marinos mantiene una tabla de protocolos taxonómicos, esta tabla puede servir de parámetro para para la ordenación jerarquizada y sistemática de la investigação científica nacional. Otra ventaja apuntada por la Unesco (2016) es que el trabajo que requiere el mantenimiento de un sistema distribuido puede ser compartido entre las organizaciones socias. También puede haber un ahorro en los costos de operación al no tener que duplicar experiencia específica.

El modelo distribuido también tiene en cuenta el volumen considerable de datos que puede ser generado por los programas de la oceanografía operacional. Estos altos volúmenes ya no pueden ser manejados por un solo centro de datos. Utilizando un enfoque de servicios web, como por ejemplo un portal de datos, el desarrollo de una red distribuida de proveedores de datos puede mejorar la capacidad de compartir e integrar los datos oceanográficos y reforzar el concepto de custodia de datos distribuidos en cada agencia responsable de proporcionar el acceso a los conjuntos de datos.

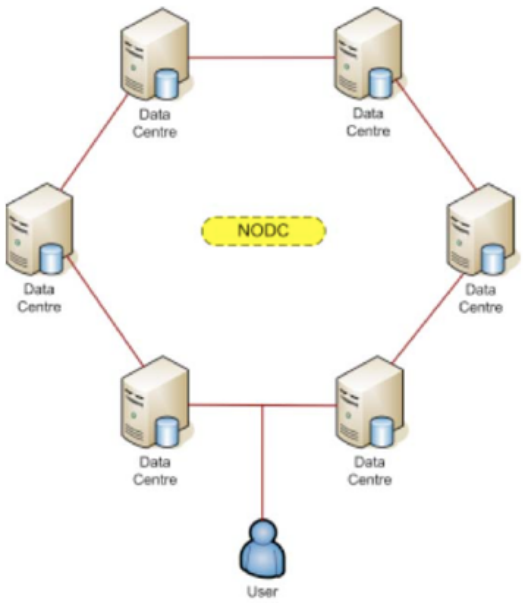

Figura 18: Modelo distribuido de centro de datos
Fuente: Intergovernmental Oceanographic Commission (2016)

En el modelo distribuido, cada proveedor de datos tiene una capacidad de gestión de datos. La posesión de datos permanece en el centro de datos de origen y se accede a los datos de forma dinámica a través de la Internet al revés de un repositorio central. Este modelo proporciona una mejor cooperación interinstitucional y la coordinación con los datos uniformes y estándares de metadatos y protocolos (COI Manuals and Guides, 2016).

Mientras que el modelo distribuido ofrece varias ventajas, también puede presentar algunas limitaciones ya que requiere:

• Alto nivel de tecnología y coordinación para configurar la red y garantizar su funcionamiento eficaz.

• La adopción de normas y protocolos aceptados para garantizar la interoperabilidad entre los diferentes componentes del sistema distribuido.

La elección de un modelo conceptual debe tener en cuenta la tecnología disponible y los recursos humanos para garantizar la eficiencia y la sostenibilidad del sistema. La extensibilidad del modelo adoptado también debe ser considerado.

3.8.3 Modelo mixto

En el modelo de centro de datos mixto, algunos proveedores de datos no tienen capacidad de gestión de datos para proporcionar datos a un centro nacional de datos oceanográficos, mientras otros presentan una capacidad de gestión de datos para vincular su centro de datos a través de una red distribuida (como puede ser em modelo de gestión de datos oceanográficos Arc Marine, que será presentado en el cap. 5).

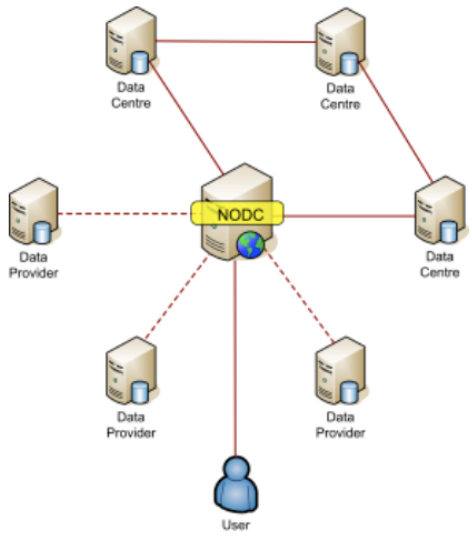

Figura 19: Modelo de centro de datos mixtos
Fuente: Intergovernmental Oceanographic Commission (2016)

Algunos proveedores crean modelos mixtos para ofrecer mayor seguridad y versatilidad para obtener y recibir datos. Una de la razones para escoger em modelo mixto es que la replicación es el método utilizado dentro del centro de datos, por lo que la mayoría de los accesos desde el almacenamiento tienen la ventaja de un rendimiento de tipo LAN. En este modelo, la codificación se utiliza a nivel local y a través de los centros distribuidores, pero una copia de todos los datos permanece en el centro de datos que más lo necesita. Luego, los datos son codificados de forma remota en todos los otros centros de datos en el ecosistema del modelo mixto.

4 SITUACIÓN INTERNACIONAL

4.1 Introducción

Con el fin de construir escenarios para el futuro y una mejor gestión de nuestros océanos, sistemas de observación del océano más precisos, completos e integrados son necesarios. Nuevas infraestructuras para observación del océano y preservación de los registros científicos han aumentado considerablemente en los últimos decenios, en la calidad, cantidad y diversidad. A medida que las nuevas tecnologías y herramientas de procesamiento de datos están remodelando nuestro conocimiento del océano, la integración de los datos de investigación en un sistema de gestión coherente y estándar sigue siendo un reto enorme en nuestra era digital.

A nivel internacional hay un permanente desarrollo de esfuerzos de los científicos e informáticos para el diseño de herramientas y métodos para explorar los océanos y entender sus interacciones críticas con el ecosistema de la tierra. Por ejemplo, tal como veremos en el presente capítulo, la Unión Europea promueve proyectos en torno a un sistema integrado de observación del océano por medio de la integración de las iniciativas de países que integran el continente. La suma de esfuerzos resulta en grandes consorcios que trabajan para ampliar el acceso y consecuente impacto de los estudios oceanográficos. Aunque no dispongan proximidad geográfica con los principales líderes en escenario mundial, en lo que se refiere a los países que poseen infraestructuras avanzadas para preservación de la investigación de los océanos, los Estados Unidos y la Australia desempeñan papel de enorme relevancia. El reto para la integración del conocimiento de los océanos esta directamente relacionado con las practicas de normalización e interoperabilidad de los datos de investigación que estés países establecen.

La interoperabilidad se refiere a un estado en el que los estándares pueden operar en más de una plataforma. Implica una norma genérica que trasciende una organización o país en particular. En las ciencias del mar, especialmente en investigaciones relacionadas el uso de redes o sistemas de sensores bajo el agua, la interoperabilidad permite un sensor conectado a un centro de

datos para ser reconocido automáticamente, por lo que el acto físico de conexión también afecta al reconocimiento del dispositivo necesario. La comunicación de datos puede entonces comenzar con el uso de estándares internacionales.

La extracción de los conocimientos sobre la dinámica y la utilización del océano son temas de actualidad, para el que recientemente han surgido nuevos desafíos tecnológicos y metodológicos. Bases de datos han crecido en escala rápidamente creciente (de los gigas a los petabytes), con frecuencias de actualización rápidos (a partir del año/mes para la escala de día/hora), e implican en estructuras complejas y de mayor dimensionalidad (de un archivo/parámetro a millones de archivos y decenas de parámetros en varios lugares). Para la comunidad oceanográfica se ha convertido en un reto para gestionar, explorar y extraer conocimiento de dichas bases de datos.

En relación a los repositorios oceanográficos, la mayoría suponen la interconexión de datos y cubren el conjunto de tareas que abarcan la investigación científica. Estos datos son imprescindibles para el entendimiento y análisis de los fenómenos que envuelven el ambiente natural marino. Son necesarios para prevenir y solucionar problemas locales, por ejemplo, las condiciones del mar, y también globales, por ejemplo, los efectos del derretimiento de las camadas polares. En tanto sea posible acceder a estos datos para análisis, mayores serán las posibilidades de encontrar soluciones preventivas para problemas ambientales y estratégicos para el desarrollo de recursos navales y científicos.

La recopilación de los datos es fruto de mediciones sobre los diferentes temas de interés expuestos anteriormente y que están empezando a ser recogidos en repositorios comunes, aunque de forma incipiente, dada a la variedad de agentes y fuentes de información que los generan (universidades, centros de investigación, agencias gubernamentales, proyectos de colaboración internacional, etc.) y su variedad estructural. Algunos de ellos se recopilan en repositorios desarrollados y bien estructurados capaces de exportar su contenido con formatos estandarizados mientras que otros se recopilan en repositorios con escasa capacidad de ser integrados con el resto, en consecuencia de la falta de uso de la

normalización internacional. En lugar de estos sitos estáticos que actúan como repositorios de datos, hay que tenerlos continuamente actualizados y asociados a funciones avanzadas analíticas y computacionales.

La misión de los centros de datos oceanográficos es facilitar el acceso y la administración de los recursos nacionales de datos oceanográficos. Este esfuerzo requiere la recopilación, control de calidad, el procesamiento, el resumen, la difusión y la preservación de los datos generados por los organismos nacionales e internacionales (IODE, 2015). Las tareas de gestión de datos que se lleva a cabo por uno centro nacional de gestión de datos oceanográficos son variadas y tienen como finalidad asegurar la conservación a largo plazo de los datos y información asociada, necesaria para la correcta interpretación de los datos.

Los consorcios de bases de datos oceanográficos están desarrollando iniciativas para mejorar el acceso a la investigación financiada con recursos públicos – y los datos relativos a las mismas – de acuerdo con normas y estándares internacionales. En estos momentos, la investigación científica es cada vez más dependiente de esos datos en la mayoría de las áreas del conocimiento.

El avance de la gestión de los datos oceanográficos puso en evidencia una revolución en el ámbito de la captura, procesamiento y análisis de los datos científicos, derivada de su volumen y complejidad. En este contexto, en algunos casos el acceso a los datos oceanográficos han encontrado apoyo en acciones consolidadas y que tienen fuerte apoyo de la comunidad científica, tales como consorcios que reúnen los datos en plataformas únicas o reunindo programas con objetivos comunes.

Desde finales de 1990 fue desarrollado una nueva estrategia de capacidades del IODE: los Ocean Data and Information Networks (ODINs). Los ODIN reúnen equipos y apoyo operacional para el desarrollo de una plataforma de red regional que puede ser utilizado por programas de la COI, como, GOOS, IODE, ICAM, tsunami, HAB, etc. Los ODIN están muy centrados en el desarrollo de los datos y productos y que suponen un enfoque de múltiples partes interesadas. También hay un fuerte enfoque en el proceso de extremo

a extremo que une observaciones, gestión de datos y el desarrollo de productos para garantizar que los centros de datos se llenan las necesidades existentes. Además, hay un enfoque en la creación de redes inter-personal e institucional. Comunicación y divulgación desempeñan un papel significativo.

Este capítulo se encuentra dividido en dos apartados principales: a) las iniciativas globales (4.2), que presenta un panorama de consorcios y proyectos en el ámbito internacional y que plantean la implementación estrategias de gestión de datos de investigación oceanográfica en nivel global, y (b) los repositorios (4.3), que presenta las principales iniciativas de países para reunir datos de investigación y ya están consolidados en el escenario internacional.

4.2 Iniciativas globales

La gestión de datos oceanográficos deriva de la necesidad para combinar, en una sola interfaz y en una escala global, la información total considerada esencial para llevar a cabo la caracterización de las condiciones ambientales que pueden esperar las fuerzas sobre los océanos. No hay compatibilidad integral entre todos los repositorios, pero mediante el uso de estándares de interoperabilidad entre los sistemas se puede lograr. Cada vez se dispone más de normas que han sido diseñados en diversas regiones del globo, pero que son aplicables al océano o datos marinos.

También existen muchos mecanismos individuales para coordinar los diferentes sistemas de datos oceanográficos. Si bien estos son esenciales para el funcionamiento continuo de la gestión de datos y el intercambio de los distintos flujos de datos, se debe poner en marcha una coordinación global para fomentar la adopción de normas, protocolos, tecnologías, etc. Las iniciativas globales deben coordinar este esfuerzo por intermedio de los consorcios que reúnen grandes repositorios de datos y desarrollar un plan de aplicación, basándose en los grupos de expertos existentes y continuando estrechos vínculos con grupos externos.

Hay muchas iniciativas que están haciendo avances en los objetivos identificados. Esto incluye el desarrollo del perfil de la comunidad marina ISO19115 para los metadatos y el trabajo en el desarrollo de vocabularios y ontologías comunes. Cada vez más se está avanzando hacia una arquitectura orientada a servicios y uso de W3C, los estándares OGC (Open Geospatial Consortium) e ISO.

El mayor desafío que se plantea en el desarrollo e implementación de la estrategia de gestión de datos e información en nivel global es una coordinación y cooperación entre los países miembros, socios y comunidades de usuarios. En este momento hay todavía importantes barreras para el uso eficiente y la reutilización de los datos. La tecnología de la información requerida para satisfacer la mayor parte de los requisitos y alcanzar una estrategia de intercambio con éxito puede desarrollarse de manera relativamente sencilla a partir de las capacidades existentes a través de la ingeniería de software. Sin embargo, la estrategia sólo tendrá éxito si todos los participantes dedican más recursos a la cooperación, utilizando activamente los datos y estándares de metadatos, protocolos de comunicación, los programas y políticas que permitan tejer las partes en un todo integrado.

A continuación presentaremos en orden alfabético las principales iniciativas globales así como los proyectos de gestión de los datos oceanográficos que tienen la finalidad de coordinar y facilitar programas de implementación para cumplir con los requisitos internacionales y mejorar la calidad de los datos:

4.2.1 ARGO Data System

ARGO es un proyecto internacional para reunir información sobre la temperatura y la salinidad de la parte superior de los océanos del mundo. ARGO utiliza una matriz global de 3.000 flotadores robóticos para medir la temperatura y la salinidad y proporcionará una descripción cuantitativa del estado evolutivo de la capa superior del océano y los patrones de la

variabilidad del clima en los océanos. ARGO tiene un equipo directivo internacional y un equipo de gestión de datos integrada por científicos de los países que participan en ARGO. (URL: http://www.argo.ucsd.edu/)

4.2.2 Coastal component of GOOS (implementado a través del GOOS Regional Alliances)

El módulo costero del GOOS contribuye con la comprensión de los efectos de la actividad humana, el cambio climático y los desastres naturales en los sistemas costeros. Coastal GOOS desarrolla un subsistema de comunicaciones y gestión de datos (DMS) para el descubrimiento y la entrega de los datos dentro de GOOS y para la interoperabilidad con otros sistemas de observación y los programas de investigación. El desarrollo del GOOS costero DMS es facilitado por la formación de un Grupo de Trabajo previsto Data Management (DMWG) en colaboración con el IODE y el Área de Programa de Gestión de Datos de la CMOMM. El DMWG formulará directrices para el desarrollo del sistema, promover el establecimiento de normas y protocolos, definir métricas y fomentar proyectos piloto. (URL: http://www.COI-goos.org/content/view/14/28/)

4.2.3 Data Buoy Cooperation Panel

El Data Buoy Cooperation Panel (DBCP) es una división del la Intergovernmental Oceanographic Commission (IOC) que presenta como principales características: revisar y analizar las necesidades de datos de boyas, coordinar y facilitar programas de implementación para cumplir con los requisitos; iniciar grupos de acción de apoyo y mejorar la cantidad y calidad de los datos de boyas distribuidas en el Sistema Mundial de Telecomunicaciones (SMT). El DBCP es administrado por el Departamento de Pesca y Océanos del Canadá. (URL: http://www.dbcp.noaa.gov/dbcp/index.html)

4.2.4 Global Climate Observing System

El Global Climate Observing System (GCOS) se estableció en 1992 para asegurar que se obtienen de las observaciones y la información necesaria para abordar las cuestiones relacionadas con el clima y la pondrán a disposición de todos los usuarios potenciales. Es co-patrocinado por la el COI. El Sistema Mundial de Observación del Clima (SMOC) está destinado a ser en largo plazo, el sistema operativo impulsado por los usuarios, capaz de asegurar las exhaustivas observaciones necesarias para el seguimiento del sistema climático, la detección y atribución del cambio climático, para la evaluación de los impactos de la variabilidad y el cambio climáticos, y para apoyar la investigación hacia la mejora de la comprensión, modelización y predicción del sistema climático. SMOC construirá, en la medida de lo posible, en los sistemas de observación y la investigación operativa, gestión de datos y distribución de información existentes, y nuevas mejoras de estos sistemas. Además, proporcionará un marco operativo para la integración y mejora, según sea necesario, de sistemas de observación de los países y organizaciones que participan en un sistema integral centrado en los requisitos para las cuestiones climáticas. (URL: http://www.wmo.ch/web/gcos/gcoshome.html)

4.2.5 Global Earth Observation System of Systems

El intergovernmental Group on Earth Observations (GEO) está liderando un esfuerzo mundial para construir un Sistema de Observación de la Tierra Global de Sistemas (GEOSS) en los próximos 10 años. El propósito de GEOSS es lograr observaciones integrales, coordinadas y sostenidas del sistema de la Tierra, con el fin de mejorar la vigilancia del estado de la Tierra, aumentar la comprensión de los procesos terrestres y mejorar la predicción del comportamiento del sistema de la Tierra. GEOSS satisfacer la necesidad de información oportuna, a largo plazo la calidad global de la información como base para la toma de decisiones acertadas, y mejorará la entrega en nueve áreas de beneficio social.

Para gestionar datos, GEOSS facilitará el desarrollo y la disponibilidad de los datos compartidos, metadatos y productos comúnmente requeridos a través

de diversas áreas de beneficio social. GEOSS alentará la adopción de las normas existentes y nuevas para apoyar los datos generales e información usabilidad. GEOSS se basará en los componentes existentes Infraestructura de Datos Espaciales (IDE) como los precedentes institucionales y técnicas en áreas tales como marcos de referencia geodésicos, datos geográficos comunes y protocolos estándar. El IOC puede beneficiarse de colaborar estrechamente con la GEOSS, en particular, con el Comité de Datos y Arquitectura medida que se desarrolla. El IOC debe estar a la vanguardia en la contribución al componente de datos marinos de GEOSS.(URL: http:// www.earthobservations.org/)

4.2.6 Global Ocean Ecosystems Dynamics

Global Ocean Ecosystems Dynamics (GLOBEC) es una respuesta internacional a la necesidad de comprender cómo el cambio global afectará la abundancia, la diversidad y la productividad de las poblaciones marinas que comprenden un componente importante de los ecosistemas oceánicos. El objetivo primordial de GLOBEC es avanzar en la comprensión de la estructura y funcionamiento del ecosistema global de los océanos, sus subsistemas principales, y su respuesta al forzamiento físico de modo que la capacidad puede ser desarrollado para predecir las respuestas del ecosistema marino al cambio global. GLOBEC utiliza un sistema de gestión de datos descentralizada, donde los metadatos se mantiene en una base de datos central y los proyectos individuales son responsables del control de calidad y archivo de sus datos. La política de datos GLOBEC se centra en los datos y el intercambio de metadatos, el archivo responsable de los datos y el inventario y las actividades de catalogación. Le da a los programas nacionales y regionales GLOBEC un marco a partir del cual se construyen sus propias políticas de gestión detallada de los datos y para abordar cuestiones tales como archivo a largo plazo de los datos para asegurar que GLOBEC hace una contribución duradera a la ciencia marina. (URL: http:// www.globec.org/)

4.2.7 Global Sea Level Observing System

El Global Sea Level Observing System (GLOSS) es un programa internacional llevado a cabo bajo los auspicios de la CMOMM para establecer redes regionales y en nivel global de monitoreo del clima, oceanografía y la investigación del nivel del mar de la costa. El GLOSS, en colaboración con el IODE, ha iniciado un proyecto destinado a la recuperación de datos de información sobre el nivel de mar disponible sólo en papel y su conversión en formato electrónico y accesible. (URL: http://www.gloss-sealevel.org/)

4.2.8 Global Temperature Salinity Profile Project

El Global Temperature Salinity Profile Project (GTSPP) es un consorcio de cooperación internacional. Su objetivo es desarrollar y mantener un recurso océano Temperatura y Salinidad mundial (TS) con datos que son posibles hasta al día y de la mejor calidad. Hacer mediciones globales de océano TS de forma rápida y de fácil acceso para los usuarios es el objetivo principal del GTSPP. Tanto los datos en tiempo real transmitidos a través del Sistema Mundial de Telecomunicación (SMT), y los datos en modo diferido recibidos por el NODC se adquieren y se incorporan en una base de datos gestionada de forma continua. (URL: http://www.nodc.noaa.gov/GTSPP/gtspp-home.html)

4.2.9 Harmful Algal Bloom Programme

El Harmful Algal Bloom Programme (HAB) busca fomentar la gestión eficaz y la investigación científica sobre las floraciones de algas nocivas a fin de comprender sus causas, predecir su ocurrencia, y mitigar sus efectos. Durante los últimos 10 años, la HAB ha establecido una serie de productos de datos que incluye: base de datos de Algas; lista taxonómica de Referencia de Algas y el directorio Internacional de algas nocivas y sus efectos sobre la pesca y la Salud Pública. (URL: http://COI.unesco.org/hab/).

4.2.10 International Council for Science

El International Council for Science (ICSU) es una organización no gubernamental que representa a una membresía global que incluye tanto los organismos científicos nacionales y uniones científicas internacionales. ICSU ha establecido una serie de organismos que se especializan en problemas de datos y de información científica a nivel internacional. El ICSU tiene un total de 52 centros en 12 países que proporcionan acceso a los datos geofísicos y ambientales para todos los científicos de forma gratuita o por el costo de la reproducción. Los centros también aseguran el archivado a largo plazo y la preservación de los datos y trabajar con los datos para mejorar su calidad. El objetivo principal ha sido la de actuar como el archivo mundial de datos, y como tal, se basan en acuerdos de intercambio de datos con los centros nacionales de datos. Hay tres ICSU para la Oceanografía ubicados en los EE.UU., Rusia y China. Además, más recientemente, un ICSU para Marine Environmental Data se ha establecido en Alemania. Además de asegurar la custodia y difusión de los datos, sino que también están en condiciones de crear o colaborar en la producción de climatologías. (URL: http:// www.ngdc.noaa.gov/wdc/)

4.2.11 International Council for the Exploration of the Sea

International Council for the Exploration of the Sea (ICES), iniciado en 1902, es la organización internacional de los tratados más antigua y la más antigua agencia oceanográfica intergubernamental. Los datos se mantienen en las áreas de la biología / de la pesca, la contaminación marina y la hidrografía clásica. ICES se refiere principalmente a la prestación de apoyo científico para la gestión internacional de los océanos. El CIES ha desarrollado una nueva estrategia de datos para hacer frente mucho más amplio, grandes conjuntos de datos en su labor futura y tiene una importante función de gestión de datos para jugar tanto como mayordomo de datos y proveedor de acceso a datos distribuidos. Tres objetivos estratégicos de gestión de datos se han identificado como (i) CIES permanecerán un punto focal para los datos marinos en el Atlántico Norte, (ii) CIES crearán un portal que sirve como un centro de datos distribuidos, y (iii) el portal web del CIES se hará más atractivo para la comunidad científica. CIES y el IOC han firmado un

memorando de entendimiento que enfatiza específicamente "la cooperación en el campo de la gestión de datos e información, incluyendo el desarrollo de tecnologías de información marinos". El IOC colabora con el CIES para asegurar que los estándares comunes están en su lugar. (URL: http:// www.ices.dk/datacentre/data_intro.asp)

4.2.12 Intergovernmental Oceanographic Commission

La Intergovernmental Oceanographic Commission (IOC) se ocupa de la gestión de datos oceanográficos internacionales, intermediado por el programa *International Oceanographic Data and Information Exchange* (IODE). La IOC fue creada para facilitar y promover el intercambio de datos y estructura distribuída para la recolección datos y para ayudar a los estados miembros en el desarrollo de las técnicas y los procedimientos necesarios para la gestión de datos oceanográficos, convirtiéndose más tarde en los socios red IODE.

En tanto que es un programa global, el IODE presta atención a todos los datos relacionados con el mar, incluida la oceanografía física, química, biológica, entre otras áreas. Por otra parte, el IODE coopera con otros programas relacionados, como la *Ocean Science, Global Ocean Observing System*, así como la *Joint Commission for Oceanography and Marine Meteorology* (JCOMM).

El programa del IODE ha establecido 65 Centros Nacionales de Datos Oceanográficos (NODC) desde su creación en 1961. A pesar de que operan a un conjunto de principios comunes, los NODC tienen una amplia variedad competencia nacional y varían en tamaño de los equipos de trabajo desde una persona a más de cien. El sistema NODC ha contribuido en gran medida a la gestión de datos oceanográficos. La mayoría de los NODC reciben datos de las agencias gubernamentales y académicas y una proporción menor de ellos (aproximadamente un tercio) también reciben los datos de las instituciones de investigación financiados con fondos privados y/o de la industria. La mayoría de los centros ofrecen datos en tiempo diferido aunque más de la mitad ofrecen datos en línea. Los NODC están manejando cada

vez más una amplia gama de tipos de datos: atmosféricos físicos, químicos y biológicos, la meteorología marina y, los datos geológicos y geofísicos y algunos manejan datos en tiempo real.

La mayoría de los NODC proporcionan datos que, con la adopción de políticas de acceso libre, se ofrecen en los repositorios que manejan grandes volúmenes. Este procedimiento permitió la creación de un portal de datos oceanográficos, que sirve para facilitar y promover el intercambio y la difusión de datos y servicios marinos a nivel mundial y a través de una red de datos nacionales y regionales federados e interoperables.

Tal como hace la *Global Earth Observing System of Systems* (GEOSS), los nuevos NODC deben construirse con sistemas e iniciativas existentes con flexibilidad suficiente para posibilitar interoperabilidad con los principales repositorios internacionales. Esta interoperabilidad debe lograrse a través del uso de estándares internacionalmente aceptados y las mejores prácticas (tales como SOA, ISO y OGC) y no requieren los centros de datos para cambiar la gestión de datos interna. El portal de datos oceanográficos fue desarrollado como un sistema que permite la interoperabilidad con los sistemas nacionales y regionales, tales como EUA IOOS, SEADATANET, *Australian Oceans Portal*, y otros, y los sistemas internacionales, como el WIS de *World Meteorological Organization* (WMO).

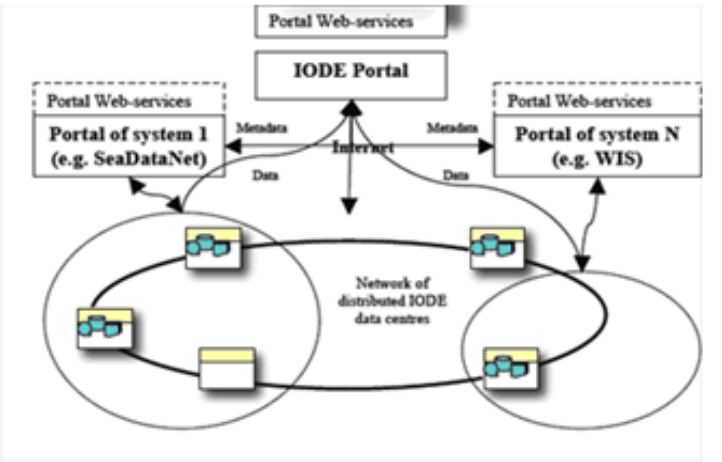

Figura 20: Flujo Portal Interoperabilidad de Servicios Web
Fuente: International Oceanographic Data and Information Exchange (2013)

El IODE busca evitar que los registros de los datos en el ámbito nacional y regional estén duplicados, como el SeaDataNet y sus extensiones regionales. Cuando no existan este tipo de redes (por ejemplo, en varias regiones en desarrollo), su estrategia consiste en proporcionar herramientas de software libre para la conexión de los sistemas de información nacionales con la red global.

Sin embargo, existe un gran reto que hay que superar para garantizar esta conexión e interoperabilidad entre los y diversos portales regionales, debido a las diferencias en las normas, las políticas y servicios, y debido al hecho de que el IODE depende de contribuciones voluntarias. Para superar las barreras técnicas es necesario que los participantes en IODE dediquen más recursos a la cooperación, a la adopción de normas comunes de metadatos, protocolos, programas y políticas para desarrollar un sistema integrado. Para ello, el IODE *y* el *Marine Meteorology and Ocean* (MMO) colaboran estrechamente con, por ejemplo a través del *Project Ocean Data Standards*, que tiene como objetivo la publicación de las normas acordadas a nivel internacional en relación con la gestión y el intercambio de datos oceanográficos.

Los NODC actúan como centralizadores nacionales para gestión de datos oceanográficos con criterios comunes para su funcionamento, aunque funcionen centralizados o dispersos en varias instalaciones, presentando proporciones diferentes en relación a capacidad para obtener, indexar y disponibilizar los datos marinos.

En la coordinación nacional hay una fusión y combinación de actividades de intercambio y archivo de datos oceanográficos, como lo indica el sistema de los NODC y sus interlocutores en materia de oceanografía operacional y la predicción del clima, la cual normalmente está vinculada a los servicios nacionales meteorológicos e hidrológicos.

La IOC no dispone una central para recogida de datos permanente que sea similar a los NODCs, de manera que posibilite prevenir las condiciones de los océanos. La falta de una central es un problema en términos de comunicación así como de organización y racionalización de las

actividades de la Comisión. Existe una laguna entre la capacidad tecnológica de los NODCs y los productos que ofrecen en cuanto a la gestión de datos y servicios de datos y productos oceanográficos operacionales. En algunos países se han tomado medidas encaminadas a dotarse de estructuras más específicas al respecto (COI, 2007).

Las características peculiares de una región determinan las prioridades en el momento de establecer criterios para gestión de datos, corroborando con la permanente necesidad de centros regionales y especializados para establecer conjuntos de datos comunes y para la distribución de productos e información. De acuerdo con la IOC (2007, p. 18), las razones que justifican la creación de centros regionales y especializados de datos e información son:

• Atender las necesidades de gestión de datos e información de una determinada Alianza Regional del GOOS;

• Atender las necesidades de una región de la IOC o de un órgano subsidiario regional;

• Atender las necesidades de datos e información de otros programas regionales, por uno de los Grandes Ecosistemas Marinos (LME) o un Programa de los Mares Regionales;

• Responder pedidos especializados, por ejemplo, de un programa científico o de un determinado servicios de datos (por ejemplo, sobre el nivel del mar);

• Afinidades regionales geopolíticas, geográficas o de otra índole (como se plantea en el enfoque de la UNESCO);

• Satisfacer las necesidades de las asociaciones regionales.

La IOC (2007) apunta que la gestión de datos oceanográficos se podría concebir que utilice sus grupos regionales (órganos subsidiarios regionales, Alianzas Regionales) como forma de organizar su enfoque regional, aplicando procedimientos similares a los de la World Meteorological Organization (WMO). Según este esquema habría una mayor responsabilidad en comparación con las circunstancias actuales, y

en esta estructura se podría dar cabida a todas las actividades de gestión de datos e información.

Conforme la IOC (2007, p. 18), los aspectos regionales de la gestión de datos e información tendrían su origen en:

• Un proyecto organizado o programa regional de importancia, con una necesidad singular de servicio regional (o especializado) de datos e información;

• Beneficios demostrables en eficiencia y/o eficacia para responder a las necesidades colectivas de una región en materia de datos e información;

• Razones importantes (necesidades) de naturaleza geográfica o geopolítica para tener un servicio regional (especializado) de datos e información oceanográficos.

En relación a los servicios de datos, un centro especializado de datos puede funcionar permanentemente o con plazo definido, dependiendo de las necesidades puntuales de su misión, con una estrategia clara para transmisión de los datos del centro especializado para un centro permanente.

Los principios de los Centros Mundiales de Datos fueron creados por el Consejo Internacional para la Ciencia (ICSU) con el objetivo de preservar los datos de las más variadas disciplinas en portales únicos, por ejemplo, Tianjin (China), Silver Spring (Estados Unidos de América), Obninsk (Federación de Rusia) y Bremen (Alemania). Los Centros Mundiales de datos Oceanográficos reciben datos e inventarios oceanográficos de manera voluntaria de los NODC del *International Oceanographic Data and Information Exchange (*IODE), de investigadores y de asociaciones de ciencias del mar (COI, 2007). Una de las metas de la Estrategia de Gestión de Datos e Información de la IOC es que haya un centro permanente de archivo a largo plazo de todos los datos, que funcione con arreglo a normas convenidas.

Según la IOC (2007, p.19), las responsabilidades mínimas de un archivo permanente consisten generalmente en que el archivo está de acuerdo en:

• Aceptar los datos y todos los metadatos de apoyo disponibles;

• Almacenarlos en su forma original o en forma tal que se puedan recuperar todos los datos y metadatos originales;

• Actualizar o modernizar el medio de almacenamiento de los datos y metadatos de modo que puedan ser leídos en el futuro;

• Suministrar datos y metadatos de apoyo a pedido de los usuarios, gratuitamente o con el costo de reproducción;

• Almacenar los conjuntos de datos de modos que puedan ser distinguidos por programa, proyecto, experimento, etc., y recuperados separadamente de datos similares.

4.2.13 International Ocean Carbon Coordination Project

El International Ocean Carbon Coordination Project (IOCCP), co-patrocinado por la IOC y el SCOR, promueve el desarrollo de una red mundial de observaciones de carbono del océano para la investigación a través de los servicios técnicos de coordinación y comunicación, acuerdos internacionales sobre normas y métodos, actividades de promoción, y enlaces a los sistemas mundiales de observación. (URL: http:// www.COI.unesco.org/COIcp/)

4.2.14 Integrated Coastal Area Management

Integrated Coastal Area Management (ICAM) es una actividad interdisciplinaria donde los científicos naturales y sociales, administradores de zonas costeras y los formuladores de políticas se centran en la forma de gestionar los diversos problemas de las zonas costeras. Los objetivos del ICAM son para hacer frente a los problemas de las zonas costeras a través de actividades de carácter más cooperativo, coordinado e interdisciplinario, y asegurar una buena coordinación entre los esfuerzos del IOC existentes relacionados con la zona costera. El programa también tiene como objetivo proporcionar un mecanismo para promover la interacción entre los programas del IOC relacionados con ICAM y los de

otras organizaciones internacionales, entre los científicos naturales marinos y científicos sociales, así como entre científicos y administradores de zonas costeras y los responsables políticos. (URL: http://COI.unesco.org/icam/)

4.2.15 Integrated Global Observing Strategy

El Integrated Global Observing Strategy (IGOS) busca proporcionar un marco general para armonizar los intereses comunes de los principales sistemas basados en el espacio e in situ para la observación global de la Tierra. IGOS es un proceso de planificación estratégica, con un número de socios que vinculan la investigación, el seguimiento a largo plazo y los programas operativos, así como los productores de datos y los usuarios, en una estructura que ayuda a determinar las brechas de observación e identificar los recursos para satisfacer las necesidades de observación. IGOS se centra principalmente en los aspectos de observación del proceso de suministro de información ambiental para la toma de decisiones y se destina a cubrir todas las formas de recogida de datos relativos a la química entorno biológico y físico humano, incluyendo los impactos asociados. IGOS ha adoptado un conjunto de datos y sistemas de información y servicios (DISS) principios y estos principios se aplican a todas las actividades de implementación de la IGOS. Los principios IGOS para datos y sistemas de información y servicios deben ser coherentes con una estrategia global e integrada que permita el uso integrado de los conjuntos de datos de múltiples fuentes. (URL: http://www.igospartners.org/)

4.2.16 International Polar Year

El International Polar Year es uno de los más grandes programas científicos internacionales con más de 50.000 científicos, técnicos, miembros de la tripulación y demás participantes de 60 naciones que participan en actividades científicas en ambas regiones polares. API cuenta con un total de 170 proyectos de coordinación con 39 de estos

proyectos ser proyectos oceanográficos, en ambas regiones polares. El flujo de datos generados por estos proyectos será grande y continuará después de que el período del API. El IOC es un firme defensor del Año Polar Internacional. La participación ha llegado al nivel del IOC organismos, programas, secretaría y los Estados miembros de gobierno. El Comité sobre IODE ha respaldado las actividades de gestión de datos para los proyectos del API oceanográficos y todos los NODC de los países activos en las regiones polares se han solicitado para coordinar las actividades con los comités nacionales del API y proporcionar asistencia para la gestión de datos oceanográficos en lo posible. API puede muy bien servir de estímulo para aumentar la conciencia sobre la importancia de la gestión de datos y una oportunidad para que los NODC de obtener recursos adicionales. (URL: http://COI.unesco.org/ipy/)

4.2.17 Joint IOC/WMO Technical Commission for Oceanography and Marine Meteorology

El Joint IOC/WMO Technical Commission for Oceanography and Marine Meteorology (JCOMM) coordina, regula y gestiona sistemas de gestión y servicios de datos que utilizan tecnologías y capacidades con tecnología de última generación, es sensible a las necesidades cambiantes de los usuarios de datos y productos del mar, e incluye un programa de divulgación para mejorar la capacidad nacional de todos los países marítimos. Trabaja en estrecha colaboración con socios como el IODE, el GOOS y el SMOC. El Área de Programa de Gestión de Datos (DMPA) implementará y mantendrá un sistema de gestión totalmente integrada de datos de extremo a extremo (E2EDM) en toda la meteorología marina y la comunidad oceanográfica. El DMPA ofrece su experiencia para ayudar a otros grupos para especificar e implementar sus propias necesidades de gestión de datos, con el objetivo general de integrar la gestión de datos en el sistema E2EDM. (URL: http://COI.unesco.org/jcomm/)

4.2.18 Large Marine Ecosystems

Large Marine Ecosystems (LME) son las regiones del océano y el espacio costero que abarcan las cuencas fluviales y los estuarios y se extienden hacia el límite marino de las plataformas continentales y los márgenes hacia el mar de los actuales sistemas costeros. Grandes ecosistemas marinos se han delineado según continuidades en sus características físicas y biológicas, incluyendo entre otras cosas: batimetría, hidrografía, productividad y poblaciones con dependencia trófica. La LME como unidad organizativa facilita las estrategias de gestión y gobierno que reconocen numerosos elementos físicos y biológicos del ecosistema y las dinámicas complejas que existen entre y entre ellos. (URL: http://www.lme.noaa.gov/Portal/)

4.2.19 Ocean Biogeographic Information System

El Ocean Biogeographic Information System (OBIS) es el programa de gestión de datos del Censo de la Vida Marina (CoML). OBIS es un proveedor basado en la web que presenta información georreferenciada global sobre las especies marinas que contienen el nivel de especie expertos y bases de datos a nivel de hábitat y ofrece una variedad de herramientas de consulta espacial para visualizar las relaciones entre éstas y su entorno. OBIS evalúa e integra los datos oceanográficos biológicos, físicos y químicos de múltiples fuentes. El Portal OBIS accede al contenido de datos, infraestructura de información y herramientas informáticas - mapas, visualizaciones y modelos - para proporcionar una instalación dinámica, global en cuatro dimensiones. (URL: http://www.iobis.org)

4.2.20 Open ocean component of GOOS

El GOOS fue implementado a través del JCOMM y es un sistema mundial permanente de observación, modelización y análisis de variables oceánicas y marinas para apoyar los servicios oceánicos operativos en todo el mundo. Existen sistemas de gestión de datos para un número de

los flujos de datos para la parte de océano abierto del GOOS. Los ejemplos incluyen la nave del programa de oportunidad (incluyendo GTSPP), boyas de datos (a través de DBCP), el nivel del mar (a través GLOSS) y ARGO.

4.2.21 Study Group on Benthic Indicators

El objetivo de este grupo de estudio es el desarrollo de indicadores sólidos de la salud bentónica. El resultado esperado será una serie de indicadores, como los marcadores geo-químicos que reflejan las condiciones biológicas, que sea fácil de usar y ampliamente aplicable en la detección de la tensión de las comunidades bentónicas. Una base de datos en línea con datos sinópticos en las comunidades macroinfaunales y las condiciones ambientales de las diferentes regiones costeras del mundo se encuentra actualmente en fase de desarrollo. (URL: http:// COI.unesco.org/benthicindicators/)

4.2.22 WMO Information System

El WMO Information System (WIS) tiene como enfoque una infraestructura global coordinada para la recopilación, la distribución, la recuperación y el acceso a los datos de la OMM y programas relacionados. El soporte que ofrece a la OMM evita incompatibilidades de datos y problemas en el intercambio de datos entre diferentes programas. Con ello se garantiza la interoperabilidad de los sistemas de información entre los programas de la OMM y fuera de la comunidad de la OMM. El CMOMM está estrechamente relacionada con el desarrollo del SIO y el DMPA ya había tomado algunas medidas que complementan la labor del SIO a través de su apoyo a la E2EDM (End to End Data Management) proyecto piloto (una actividad conjunta IODE-CMOMM).(URL: http://www.wmo.ch/web/www/WISweb/ home.html)

4.2.23 World Climate Research Programme

El World Climate Research Programme (WCRP) está patrocinado por el CIUC, la OMM y la IOC. Los dos objetivos principales del PMIC son determinar la predictibilidad del clima; y determinar el efecto de las actividades humanas sobre el clima. El PMIC abarca estudios de la atmósfera global, océanos, mar y hielo terrestre, la biosfera y la superficie de la tierra. PMIC ha establecido un grupo de trabajo sobre la gestión de datos para desarrollar actividades comunes de gestión de datos, para asegurar la disponibilidad de los datos para la asimilación y el desarrollo de nuevas técnicas de asimilación. (URL http://wcrp.wmo.int/)

4.2.24 Working Group on Coral Bleaching and Local Ecological Responses

El objetivo de este grupo es integrar, sintetizar y desarrollar la investigación mundial sobre la decoloración de los corales y los impactos ecológicos relacionados con el cambio climático en los ecosistemas de coral, y otros nuevos hallazgos de la investigación en el desarrollo de herramientas y técnicas para la mejora de las observaciones, predicciones y las intervenciones de gestión a nivel nacional y escalas globales. (URL: http://COI.unesco.org/coralbleaching/)

4.3 Repositorios

El establecimiento de repositorios de datos integran los diferentes centros e instituciones relevantes en investigación marina. Estos repositorios están organizados de forma que se puedan atender los compromisos nacionales en redes de mayor ámbito geográfico o de cualquier otro tipo.

Los repositorios de datos oceanográficos fueron creados con el objetivo de constituir un repositorio colectivo que fuera un instrumento de consulta para datos de todos los centros de investigación de un único país, pero rápidamente ampliaran sus actividades. Con la creciente necesidad de intercambio de datos, pasaran a reunir los repositorios de centros nacionales de datos de otros

países, así constituyendo una red colaborativa, diseñadas e implementadas para que fuera posible la interoperabilidad de datos entre ellos.

La interoperabilidad se basa en las normas elaboradas a nivel internacional, que incluye especificaciones de interoperabilidad de la Comunidad considerados como estándares. La mayoría de las normas de estos repositorios de datos han sido aplicadas por las soluciones de la comunidad científica oceanográfica. Actualmente incluyen repositorios electrónicos cooperativos que contiene documentos de investigación, incluyendo imágenes, audio, datos observaciones en general y todo tipo de material relacionado con investigaciones marina.

Los repositorios tienen diversas características comunes: cumplen con protocolos de interoperabilidad y metadatos; crean mecanismos de estadísticas del acceso a los datos; proporcionan instrumentos de visualización y búsqueda conjunta de documentos que fomente el uso de los mismos; están construidos y alimentados de manera cooperativa. Estos repositorios están instalados en un clúster de alta disponibilidad con las características de balanceo de carga de las consultas que reciben, y de tolerancia a fallos en caso de desastre en alguno de los nodos que componen la plataforma.

A continuación se especifican, en orden alfabético, los principales repositorios internacionales, las principales características de cada uno y el ámbito geográfico en el que actúan.

4.3.1 Australian Ocean Data Center Joint Facility (AODCJF)

Australia, al igual que muchos países, ha tenido durante mucho tiempo un centro de datos oceanográficos nacional. Sin embargo, la tendencia mundial en los últimos años había sido la de pasar del concepto de un solo centro a una distribución, a modo de red de la operación. Por eso, en 2005 se creó el Australian Ocean Data Centre Joint Facility (AODCJF) para proporcionar la gestión de datos por el gobierno Australiano[24].

[24] Es importante tener en cuenta que al hacer esta transición, el enfoque se amplió deliberadamente del termo 'oceanográfica' (lo que implica de acuerdo con la AODC-JF, un enfoque en datos físicos) para "océano" (que abarca todos los tipos de datos marinos).

El AODN proporciona las herramientas y servicios necesarios para los flujos de datos de extremo a extremo y también tiene la capacidad de crear "puntos de vista regionales del sistema nacional, es decir, para permitir una jurisdicción particular para gestionar sus datos oceánicos sin tener que crear la infraestructura de información independiente.

El AODCJF consiste en la unión de las seis agencias del gobierno federal, lo que representa un sistema de gestión de red distribuida de datos oceanográficos. Ellos son: *Australian Institute of Marine Science* (AIMS), *Australian Antarctic* Division (AAD), *Bureau of Meteorology* (BOM), *Commonwealth Scientific and Industrial Research Organisation Marine and Atmospheric Research* (CMAR), *Geoscience Australia* (GA) y del Departamento de Defensa (Marina Real Australiana). El AODCJF actúa en proyectos que ven el desarrollo de un sistema distribuido y denominado *Australian Ocean Data Network* (AODN). Este a su vez es responsable de la interoperabilidad de datos con la comunidad marina.

Por más de 40 años, el AIMS ha jugado un papel fundamental en el suministro de esta información para aguas tropicales del norte de Australia. El esfuerzo de investigación de la AIMS está diseñado para cumplir con los desafíos que enfrentan las investigaciones en los ecosistemas marinos y avanzos económicos que dependen del mar. Pesca, petróleo y gas, la minería, el turismo y la acuicultura arrecife todos se han beneficiado de la investigación orientada a la AIMS la protección y el desarrollo sostenible de los recursos marinos. Más de cuatro décadas ha establecido avanzados programas de vigilancia, repositorios de datos e inteligencia ambiental que permite AIMS cuantificar cambios en el sistema para entender el contexto de la biodiversidad marina y la economía marítima. La AIMS é una autoridad mundial en las investigaciones marinas, logrando reconocimiento internacional por el desarrollo de su trabajo. La amplia experiencia y la fuerza de sus relaciones de colaboración nacionales e internacionales impulsan un enfoque multidisciplinar a gran escala y largo plazo. El Instituto mantiene instalaciones especializadas de investigación marina en apoyo a sus objetivos; entre ellas se encuentran el acuario Sea inteligente

Simulador Nacional[25], y una flota de buques que apoyen el acceso a los ecosistemas a través de la plataforma continental y cerca de la orilla. Direcciones documento estratégico del Instituto muestra cómo AIMS mariscales sus fortalezas y recursos para lograr sus metas.

El AIMS tiene la supervisión de la Australian National Data Service (ANDS)[26], la cual actúa como coordinador de la gestión de datos de la investigación australianos. Se trata de un proyecto nacional orientado para que los investigadores cuenten con datos de alta calidad que puedan ser reutilizables. Para ello es necesario que los datos pasen de una situación en la que son inmanejables, invisibles, están desconectados y son de uso particular a convertirse en colecciones de datos estructurados, manejables, conectados, y que pueden ser encontrados y reutilizados.

El ANDS reduce el proceso de registro de datos por medio de la construcción de infraestructuras que permiten la integración de todos los procesos relacionados con la creación o captura de datos. El avance tecnológico para gestión de datos en Australia está directamente relacionado con el desarrollo de softwares por la ANDS destinados a una correcta descripción de los datos de investigación y infraestructuras e instalaciones de almacenamiento de metadatos. Esta integración de procesos facilita a los investigadores compartir los datos mediante el Australian Research Data[27], promovendo la visibilidad de las colecciones de datos de investigación de los investigadores en los motores de búsqueda.

Australia Occidental es el primer nodo regional de la AODN, con excelentes capacidades de computación y las iniciativas de tecnología marinas por satélite son reunidas bajo un mismo centro. A pesar de la naturaleza distribuida de fuentes de datos, la comunidad marina de Australia Occidental se caracteriza por una estructura de gestión de datos fuertemente centralizada que hace que el acceso a datos por el AODN sea mucho más factible.

---

25 http://simulationaustralia.org.au

26 www.ands.org.au

27 http://researchdata.ands.org.au

4.3.2 Banco Nacional de Dados Oceanográficos (BNDO)

La Marina de Brasil es la Institución Nacional, cuyas funciones son promover y coordinar la participación del país en las actividades de la COI relacionadas con los servicios del Océano y la cartografía oceánica, sirviendo como Banco Nacional de Datos Oceanográficos. Además, la Marina de Brasil es la depositaria de la COI con la finalidad de integrar el sistema global para datos oceanográficos.

El Centro de Hidrografia Naval (CHM), Organización Militar de la Marina de Brasil, es responsable de operar el BNDO a través de la Superintendencia de Información Ambiental, cuyas actividades son:

- Obtener, recibir, analizar, verificar la consistencia de los datos recibidos, gestionar y difundir datos oceanográficos;

- Mantener interncâmbio datos oceanográficos con redes en Brasil y congéneres instituciones extranjeras dentro de la COI;

- Mantener la colección bibliográfica de las publicaciones y documentos de la COI, para su difusión entre la comunidad científica nacional;

coordinar, vigilar y supervisar, con la participación del Ministerio de Ciencia y Tecnología, la vinculación con programas de datos oceanográficos.

4.3.3 British Oceanographic Data Centre (BODC)

El BODC es un centro de datos a cargo de los datos marinos británicos biológicos, químicos, físicos y geofísicos adquiridos por diversos equipos y de diferentes fuentes, divididas en tres bases de datos diferentes:

- National Oceanographic Database (NODB);
- Project Database;
- Web Database.

Las bases de datos BODC fueron diseñadas e implementadas con el programa Oracle. Su infraestructura permite la migración de archivos de datos en diferentes formatos, la adaptación de los datos originales de las normas adoptadas por la BODC. Sin embargo, los archivos originales también se

almacenan como una medida de seguridad. Cuando no hay datos específicos sobre los parámetros individuales realizados en una base de datos de la encuesta dados, se requiere llenar un formulario indicando que el método y algoritmos utilizados para realizar los cálculos. En el portal BODC existen tutoriales sobre los procedimientos necesarios para enviar datos para identificar posibles normas adoptadas formato de archivo y presentado.

4.3.4 Centro Nacional de Datos Oceanográficos (CeNDO) de México

Sus principales objetivos son estandarizar, sistematizar e implementar una base de datos oceanográficos nacionales para la administración, la integración y el intercambio rápido entre diferentes instituciones y niveles de gobierno y los diferentes sectores de la sociedad y de las organizaciones internacionales con interés en las actividades marinas. El CeNDO también funciona como un repositorio de datos e información ambientales del Sistema Nacional de Monitoreo Oceanográfico del México (SINAMO), un instrumento desarrollado por la Comisión Interministerial para la Ordenación Sostenible de los Océanos y las Zonas Costeras (CIMARES). También participa en la cooperación internacional para el intercambio de datos e información oceanográficos ser el punto focal para la *International Oceanographic Data and Information Exchange* (IODE), vinculada à Comissão *Intergovernmental Oceanographic Commission* (COI), perteneciente a la UNESCO.

4.3.4 Centro Argentino de Datos Oceanográficos (CEADO)

Mantiene bases de datos para la explotación y desarrollo de la ciencia del mar. Entre los servicios que ofrece están:

- Los datos físicos y químicos (Atlántico SW);
- Los datos físicos y químicos (Océano Austral),
- Los datos de temperatura (estaciones base);
- Investigaciones oceanográficas publicadas;
- Las publicaciones de la Comisión Oceanográfica Intergubernamental;

- Intermediación de las organizaciones vinculadas a las Ciencias Nacional Oceánico, Grupo Argentino de Programas Científicos da la UNESCO (GAPCU) y el Programa Internacional de Boyas del Atlántico Sur (ISABP).

4.3.5 European Marine Observation and Data Network (EMODnet)

Es una iniciativa de la Unión Europea (UE) en calidad de desarrollo de la red de sistemas de observación europeos, unidos por un marco de gestión de datos que abarque todas las aguas costeras de los mares europeos y las cuencas oceánicas adyacentes. Apoya y proporciona los datos de WISE-Marine un componente marino de la Agencia Europea de Medio Ambiente de la *Shared Environmental Information System* (SEIS). La infraestructura y los estándares de SeaDataNet fueron adoptados para la gestión de datos, incluyendo los químicos, físicos e hidrográficos.

4.3.6 Geo-Seas

Está implementando una infraestructura de datos geológicos y geofísicos marinos procedentes de 26 centros de datos europeos. Esta combinación de datos se logra a través de la adopción de tecnologías y metodologías desarrolladas por el SeaDataNet en su proyecto. Sus planes de ejecución se ajustan a las directivas europeas y los programas a gran escala de los últimos en la elaboración escala mundial, como la *Global Earth Observing System of Systems* (GEOSS), GMES y EMODnet.

4.3.7 Integrated Marine Observing System (IMOS)

El IMOS es un conjunto de equipos establecido y mantenido en el mar, proporcionando flujos de datos oceanográficos *in situ* y servicios de información que, en conjunto contribuyen para atender las necesidades de la investigación marina en ambos océanos abiertos y océanos costeros alrededor de Australia, gestionado a nivel nacional y distribuido. Combinado con los datos de satélite, que proporciona esencial en datos in situ para comprender y modelar el papel de los océanos en el cambio climático, y los datos para inicializar los modelos

de predicción climática estacional. Si se mantienen en el largo plazo, permitirá la identificación y la gestión del cambio climático en el medio marino costero. Además, proporcionará un nexo de observación para entender mejor y predecir conexiones fundamentales entre los procesos biológicos y fenómenos costeros regionales / oceánicas que influyen en la biodiversidad. Mientras que como un proyecto NCRIS IMOS fue diseñado para apoyar la investigación, los flujos de datos también son útiles para muchos de la sociedad, el medio ambiente y las aplicaciones económicas, como la gestión de los recursos naturales marinos y sus ecosistemas asociados, apoyo y gestión de industrias costeras y en alta mar, la seguridad marítima, el turismo marino y defensa.

El IMOS fue diseñado para recopilar datos de las cuencas oceánicas observación y escalas regionales, incluidos los datos físicos, químicos y variables biológicas. Está operado por diez instituciones diferentes dentro del Sistema Nacional de Innovación, con el financiamiento para implementar equipos y proporcionar los datos de la demanda Australian Navy y toda la comunidad de investigación científica del cambio climático y sus colaboradores internacionales. El portal de datos del IMOS atiende investigadores interesados en la exploración de los flujos de datos de las instalaciones de recogida de datos marinos, algunos casi en tiempo real. Los datos de sus instalaciones son monitoreadas en colaboración entre la marina australiana y de la comunidad científica del clima. Se trata de una comunidad extensa y diversa desarrollada a través de una serie de nodos integrados con un Bluewater, un "nodo centrado en mar abierto," y cinco nodos regionales, incluyendo la plataforma costera y los mares Australia Occidental, Queensland, Nueva Gales del Sur, Australia del Sur y Tasmania. Los responsables de la investigación de los “links” se unen para formar una comisión nacional que supervisa todo el proceso.

La gobernanza del IMOS es controlada por un Consejo Asesor y tiene un Presidente independiente. Los miembros del Consejo son nombrados según la área que actúan para guiar el programa y ofrecer una perspectiva integradora sobre el desarrollo del IMOS. La Oficina está sediada en la Universidad de Tasmania la cual coordina y gestiona todas las inversiones como un sistema nacional.

La distribución científica del IMOS es fijada por cinco nodos regionales que cubren la Gran Barrera de Coral, Nueva Gales del Sur (sureste de Australia), el sur de Australia, Australia Occidental, las extensiones de la geografía marítima australiana y el y análisis del clima. Cada nodo tiene entre 50 y 100 miembros. Bajo la dirección central del IMOS nueve salones nacionales hacen las observaciones especificadas por los nodos utilizando diferentes componentes de la infraestructura y los instrumentos, por ejemplo, hay instalaciones separadas para los flotadores Argo, radar costero, etc. Existen tres instalaciones de observación para las extensiones de la geografía marítima australiana y observaciones climáticas (Argo Australia, mediciones mejoradas por los buques de oportunidades y del Southern Ocean Series de Tiempo), tres instalaciones de las corrientes costeras y las propiedades del agua *coastal currents and water properties* (Amarres, Ocean Planeadores y HF Radar) y tres para los ecosistemas costeros *coastal ecosystems* (Acústica de marcado y rastreo, Autonomous Underwater Vehicle y una red de sensores biofísica en la Gran Barrera de Coral).

Los operadores de las instalaciones son los principales actores en la investigación marina en Australia. Un satélite con instalaciones de teledetección reúne datos para la región y la Infraestructura de la Información electrónica Marina proporciona acceso a todos los datos IMOS y servicios web por medio de búsqueda interoperable.

El valor de esta inversión en infraestructura se encuentra en el despliegue coordinado a nivel nacional de una amplia gama de equipamiento destinado a derivar conjuntos de datos críticos dentro de una región que sirven para múltiples aplicaciones. La ejecución de instalaciones IMOS se inició en 2007, y más del 90% de la infraestructura planeada ahora se ha desplegado. El acceso libre y gratuito a los flujos de datos IMOS está disponible casi en tiempo real a través de su portal web.

Australia está en la vanguardia de los avances internacionales de la predicción y el análisis operativo del océano y el análisis del clima (a través del Australian Community Climate and Earth System Simulator (ACCESS). El desarrollo y la integración continua de sus cuatro Nodos Regionales se basa en el objetivo estratégico IMOS para el seguimiento coherente de corrientes limítrofes de

Australia. Además de lograr un equilibrio adecuado a través de escalas de espacio y tiempo, el IMOS tiene el reto de tomar decisiones sabias acerca de qué variables debe observar y cómo. Esta estrategia es bien definida para la física, la biología y la maduración de la biogeoquímica.

4.3.8 JERICO

Se trata de una infraestructura de red de investigación europea común para los observatorios costeros. Tiene como objetivo desarrollar una mejor coordinación entre los observatorios costeros responsables de los parámetros físicos y bioquímicos de los océanos, así como contribuir al desarrollo de nuevos sensores, procedimientos, estrategias de control la calidad y el cambio de *know-how* en la instalación y el funcionamiento de observatorios costeros. Las normas SeaDataNet y los servicios están siendo adoptadas por el proyecto JERICO para la gestión del flujo de datos en tiempo real procedentes de sensores para la previsión MyOcean y de los servicios y los centros SeaDataNet de datos.

La SeaDataNet estableció una estrecha cooperación con las bases de datos EuroGOOS, COI/IODE y el consorcio MyOcean**.** La EuroGOOS es una asociación de organizaciones gubernamentales nacionales y organizaciones de investigación comprometidas con el nivel europeo en el contexto de la oceanografía operacional System Global Ocean Observing Intergovernmental (SGOOI). MyOcean integra el despliegue de vigilancia paneuropeo de los océanos, especialmente actuando en la gestión de datos armonizados sobre la oceanografía física y el apoyo a la oceanografía operacional.

Todos estos desarrollos realizados por la comunidad que utiliza SeaDataNet, en cooperación con las comunidades asociadas que participan en proyectos financiados por la UE relacionados con Geo-Mares y EUROFLEETS, presentan normas para formatos de metadatos, datos y productos de datos, la calidad, los métodos de control de calidad y vocabularios comunes y también como resultado de los servicios de recogida de datos, visualización y descarga. Además ofrecen las herramientas de software para la edición de los datos, la conversión, el análisis, la comunicación y la presentación que están adoptando actualmente por la comunidad europea y los estudios marinos. La amplia gama

de tipos de datos actualmente en uso en el medio marino, así como la adopción y adaptación de los próximos nuevos estándares básicos de la Organización Internacional de Normalización (ISO), que define la información necesaria para documentar los datos y OGC proporciona una serie posibilidades de gestión de datos de calidad sobre los océanos y los estándares marinos. En este contexto, SeaDataNet II y los proyectos relacionados actúan en la explotación y el desarrollo de normas adicionales y nuevos.

4.3.9 MyOcean

MyOcean es una serie de proyectos concedidos por la Comisión Europea dentro del programa Global Monitoring for Environment and Security (GMES), cuyo objetivo es desarrollar (definición, diseño, desarrollo y validación) la capacidad paneuropea para la monitorización y predicción oceánica. Las actividades se benefician de varias áreas específicas de uso: seguridad marítima, prevención de derrames de petróleo, gestión de los recursos marinos, el cambio climático, las actividades costeras y la calidad del agua. La serie de proyectos MyOcean terminó en 2015, y sus servicios ahora está continuó por el Programa de Copérnico.

4.3.10 National Oceanic and Atmospheric Administration (NOAA)

Pertenecientes al Departamento de Comercio, es una empresa líder en el desarrollo e implementación de la *Integrated Ocean Observing System* (IOOS, EEUU)**.** Se nutre principalmente de contribuciones de los gobiernos y las investigaciones independientes que desarrollan estudios marinos, la creación de un sistema de forma rápida y sistemática de adquirir y difundir datos sobre el mar, la costa y los lagos.

Un factor esencial para el éxito de IOOS es la presencia del *Data Management and Communication System* (DMAC), capaz de proporcionar datos en tiempo real para las observaciones de teledetección de física, química y biológica.

4.3.11 National Oceanographic Data Center (NODC)

El NODC de los EEUU es uno de los centros de datos nacionales del medio ambiente operados por la NOAA. Se ocupa del mantenimiento y la actualización de los archivos de datos medioambientales, organización de datos para ayudar a monitorear los cambios ambientales globales, como las mediciones físicas, químicas y de investigación oceanográfica biológica obtenida mediante la teleobservación por satélite de los océanos. Para obtener los datos, el NODC interactúa directamente con las autoridades federales, estatales, instituciones académicas e industriales que trabajan en actividades oceanográficas, en representación de NOAA en varios paneles, entre los comités y consejos nacionales, además de representar a los EEUU en los organismos internacionales, como la *International Oceanographic Data and Information* Exchange (IODE), y la *Intergovernmental Oceanographic Commission* (COI).

4.3.12 Ocean Data Portal (ODP)

El objetivo del Ocean Data Portal (ODP) es promover el intercambio de datos y servicios de investigación marina. El ODP proporciona acceso a las colecciones y los inventarios de datos marinos de los NODC en la red IODE y posibilita la evaluación (a través de la visualización y revisión de los metadatos) y el acceso a datos a través de servicios web. La arquitectura del sistema utiliza tecnologías orientadas para acceder a datos e información oceanográfica no homogéneas y geográficamente distribuidas.

El ODP se desarrolla en estrecha cooperación con iniciativas como el IODE, SEADATANET y otros. Además, apoya los requisitos de acceso a datos de otras áreas de programas de la COI, entre ellos el GOOS y el sistema de alerta de tsunami. El ODP también trabaja en estrecha colaboración con otras iniciativas internacionales, incluyendo el GEOSS para garantizar la interoperabilidad con otros dominios.

4.3.13 Rolling Deck to Repository (R2R)

En esencia, el R2R se utiliza para recopilar datos de expediciones oceanográficas habitualmente catalogados, y posteriormente compartido con bases de datos nacionales, incluyendo el Centro Nacional de Datos Geofísicos

(NGDC) y NODC. El programa R2R define los datos digitales generados por los sistemas de sensores instalados de forma permanente a bordo de la nave y de manera rutinaria mantenido por el operador, incluyendo:

Metadatos	*Descripción / Ejemplos*
ADCP	Acústico Doppler Current Profiler
CTD	Conductividad, temperatura, presión, y otros sensores en la columna de agua
Sonda	Profundidad acústica de frecuencia única o múltiple de los fondos marinos o de media agua reflectores
Sonda dispensável	XBT, XCTD, XSV, e outros
Fluorometer	Fluorescencia (normalmente para el fitoplancton)
Gravímetro	Campo de Gravedad
Magnetómetro	Campo magnético
Estación Meteorológica	Datos meteorológicos (viento, por ejemplo, la temperatura, la humedad, la turbulencia, humedad)
Multifeixe	Montado en el casco de imagen los datos de asignación de sonar del fondo marino
Navegación	Posición y movimiento como el Sistema de Posicionamiento Global (D / GPS, WAAS), Unidad de Referencia Vertical (IMU / MRU), récord de velocidad, Girocompás
pCO2	La presión parcial de dióxido de carbono disuelto
SSV	La superficie del mar de sonido del velocímetro
Subbottom	Acústicas de los sedimentos de penetración de perfiles de datos de superficie
TSG	Thermosalinograph - fluxo de dados através de temperatura e salinidade
Guincho	La tensión del hilo, la velocidad de pago, etc

Tabla 13: Metadatos utilizado por el programa Rolling Deck to Repository
Fuente: Rolling Deck to Repository (2013)

El proyecto R2R está supervisado conjuntamente por un equipo de la Universidad de California –*San Diego Supercomputer Center* (SDSC), el Instituto Scripps de Oceanografía– *Geological Centro de Datos* (GDC-SOI), el *Lamont Doherty Earth Observatory* (LDEO) de la Universidad de Columbia, el Instituto Oceanográfico *Woods Hole* (WHOI) y la Universidad Estatal de Florida (FSU). Los datos recogidos en cada una de estas expediciones oceanográficas catalogadas se transmiten de forma rutinaria a los portales nacionales de Estados Unidos, incluyendo la National Geophysical Data Center (USGS) y el

National Oceanografic Data Center (NODC). La recogida se lleva a cabo mediante el uso de herramientas específicas para capturar y enviar los datos de las rutas de los buques de forma automática, lo que garantiza la conservación de datos de rutina en curso. Durante los tres años de su existencia, el R2R ha capturado datos sobre más de 3.000 cruceros, con 25 barcos y con un total de más de 13 millones de archivos de datos (11 TB). R2R está desarrollando nuevos modelos para el archivo de datos con el fin de organizar la variedad de las actividades a bordo de buques, por lo general se realizan durante expediciones en buques modernos.

4.3.14 SEADATANET

Esta es la principal red europea operativa activa, constituye la infraestructura para la gestión, indexación y el acceso a una amplia gama de datos marinos adquiridos por cruceros de investigación y actividades de observación en aguas y océanos de todo el mundo. Conecta los *National Oceanographic Data Centres* (NODC) y los servicios de información de los grandes centros de investigación marina de 35 estados costeros que bordean los mares europeos.

Hay varios centros de datos que integran SeaDataNet y que están por toda Europa. Se encuentra en los institutos de investigación marina, con capacidad para la gestión de datos, recuperación y distribución de datos oceanográficos. Se mantienen las redes nacionales con la participación de otras organizaciones en diferentes países que ejercen la autonomía para la gestión de datos marinos y oceanográficos. En conjunto, las actividades de estos centros de datos se extienden desde la costa hasta las profundidades del océano, incluyendo actividades de investigación marina y vigilancia del medio ambiente a través de una variedad de temas, tales como el cambio climático, la hidrodinámica, la geología los recursos vivos marinos, la biodiversidad y los hábitats.

El SeaDataNet establece normas para indexación de los datos marinos que están adaptadas para muchos otros proyectos en Europa. El objetivo principal es ofrecer los metadatos marinos derivados de la investigación llevada a cabo por los estados costeros europeos implicados en el proyecto. De esta manera, ha adoptado una política común de datos, ofrece la búsqueda de forma gratuita,

aunque el acceso sea controlado a todos conjuntos de datos. Los costos son controlados por una plataforma de datos distribuidos, que están enlazados y accesibles para los usuarios a través de un portal central. El proyecto Geo-Seas de la geología y la geofísica y la actualización *Upgrade Black Sea SCENE* (UBSs), dedicado a la región del Mar Negro son dos ejemplos de proyectos que han adoptado las normas SeaDataNet.

4.3.15 Systèmes d'Informations Scientifiques pour la MER (SISMER)

El SISMER es un centro de datos oceanográfico francés, bajo la responsabilidad del Instituto Francés de Investigación para la Explotación del Mar, llamado Institut Français de Recherche pour l'explotacion de la Mer (IFREMER), que actúa en la gestión de física, química y geofísica. Sus datos y metadatos se distribuyen en tres fuentes, de la siguiente manera:

• DATABASES directory: es una lista organizada de todos los datos marinos recogidos por la comunidad científica francesa y bases de datos internacionales, en representación de las investigaciones de Francia al Directorio Europeo de Datos Ambientales Marinos (EDMED);

• OCEANOGRAPHIC CRUISES directory: cumple informes de cruceros para la *Report of Observations/Samples collected by Oceanographic Programmes* (ROSCOP) de los barcos de investigación de Francia y asociaciones desarrollado para la cooperación mutua con instituciones extranjeras;

• Data archived at SISMER: el repositorio de datos del SISMER.

Tanto el directorio de base de datos como los cruceros oceanográficos del directorio proporcionan sólo metadatos, mientras que los datos archivados en SISMER proporcionan datos y observaciones más específicas.

Su uso es público, pero con el derecho a la privacidad y la exclusividad otorgados por el IFREMER, presentando diferentes formatos, tales como:

- NetCDF multiprofile;
- NetCDF grid;
- ASCII multiprofile;

- GTSPP (Global Temperature & Salinity Profile Project).

El formato NetCDF o ASCII MultiProfile presentan una organización interna particular en un solo archivo son todas las encuestas en un crucero para un tipo particular de equipo. Ya formato GTSPP fue adoptada por el mismo nombre.

4.4 Evaluación de repositórios

Las redes internacionales de sistemas de gestión de datos integran la observación oceanográfica y la gestión de la comunicación marina para la recepción de datos, estructurados por un conjunto de herramientas dirigidas a los usuarios.

Una barrera importante para la investigación geocientífica marina es la falta de organización de datos geofísicos y productos de datos y, aunque haya gran volumen de datos geológicos y geofísicos disponibles para el medio marino, a menudo es muy difícil de utilizar estos conjuntos de datos de una manera integrada. Esto es en parte debido a la utilización de nomenclaturas diferentes, formatos, escalas y sistemas de coordenadas, tanto entre diferentes organizaciones, así como a través de las fronteras nacionales. Esto hace que sea muy difícil el uso directo de los datos de geociencias marinas primarias y dificulta la utilización de estos datos para producir los productos de datos multidisciplinares integrados y servicios.

La información geológica y geofísica marina incluye datos observacionales, datos brutos y datos analíticos, estudios geofísicos (sísmica, gravedad, etc), multihaz y estudios de sonar de barrido lateral, así como productos de datos derivados, como los mapas del fondo marino. Todos los cuales son necesarios con el fin de producir una interpretación geológica completa del fondo marino por intermedio de una IDE (Infraestructura de Datos Espaciales), o sea, una arquitectura integrada por un conjunto de recursos (catálogos, servidores, programas, datos, aplicaciones, etc) que sirven para gestionar Información Geográfica (mapas, fotos, imágenes de satélite, topónimos y otros). Esos recursos desempeñan variados entornos de interoperabilidad (especificaciones, interfaces, protocolos, normas y otros) y admiten que un usuario pueda

utilizarlos y combinarlos según sus necesidades. Así, el objetivo de una IDE es Integrar a través de Internet los datos, metadatos, servicios e información de tipo geográfico que se producen en determinada región, facilitando a todos los usuarios potenciales la localización, identificación, selección y acceso a tales recursos.

Actualmente el desarrollo de investigaciones en las áreas oceanográfica y geoespacial presentan relaciones muy próximas sobre la recogida de datos, aunque el almacenamiento, configuración y análisis sean tratadas de manera distinta para representar un mismo fenómeno. En cuanto los datos geoespaciales abarcan dos a tres medidas dimensionales, los datos oceanográficos comprenden diversas cuantificaciones multidimensionales, tales como corrientes marinas, salinidad, atmósfera, mapas de pesca y muchos otros. Estas diferencias de modelos para gestión de datos resultan en dificultades para ambas comunidades presentaren y analizar los conjuntos de datos que disponen de forma integrada.

Para realizar el análisis de la situación internacional, evaluamos cuál es el grado de desarrollo de los repositorios de datos científicos en oceanografía y, además, un análisis de las características de estos repositorios a partir de un conjunto de indicadores. Se dispone de algunos estudios parciales como el de Sydney Levitus (Levitus, Antonov, Baranova, Boyer, Coleman, Garcia, et al., 2013), que analizan el funcionamento de la World Ocean Database (WOD), una de las agencias de investigación que componen la NOAA, pero no se encuentran estudios globales como los que planteamos.

Se ha llevado a cabo una evaluación a partir del análisis de los propios repositorios en base a un conjunto de indicadores establecidos, algunos de ellos difíciles de evaluar mediante consulta externa a los repositorios. No se trata de un cuestionario enviado a los responsables de los repositorios.

Para realizar análisis de los repositorios de datos oceanográficos, se ha interrogado las bases de datos de la International Ocean Discovery Program (IODP), Woods Hole Oceanographic, Ocean Docs, Aquatic Commons y Seadatanet. Los resultados apuntaran bases de datos nacionales y directorios que reúnen un conjunto de otras bases.

También se ha realizado análisis de los repositorios de datos oceanográficos consultando repositorios de acceso abierto combinando las palabras clave: *ocean data*, *oceanographic data* y *repository oceanographic.* Las búsquedas han sido realizadas en las siguientes bases: ODISEA (International Registry on Research Data), ROAR (Registry of Open Access Repositories), OpenDoar (Directory of Open Access Repositories), y también en re3data.org.

No se incluyen proyectos de recuperación de datos oceanográficos como puede ser UNESCO-COI-IODE, que tiene el objetivo de localizar y digitalizar los datos existentes en forma de manuscrito o electrónico que se encuentran en riesgo de pérdida debido a la descomposición (Levitus, 2012).

4.4.1 Objetivos y metodología

El análisis de los repositorios tuvo como objetivo identificar los diferentes formatos de registro, los sistemas de difusión de datos, las tecnologías utilizadas para la carga y la integración de datos oceanográficos, entre otros.

Para el análisis se han establecido los siguientes indicadores:

Indicadores	Descripción	Valores
Ámbito geográfico	Cobertura geográfica donde los datos fueran recogidos.	Nombre del país o del continente.
Institución	Indicación de los organismos responsables de la recogida de datos.	Nombre de la institución.
Tipo de servicio	Se refiere a la estructura y funcionamiento del repositorio, ya se trate de un proveedor de datos (incorpora directamente los datos) o agregador (recolecta los datos de otros repositorios).	Proveedor, agregador
Tipología de datos	Indicación de las características de los datos recogidos.	Físicos, químicos, geológicos, biológicos.
Formato de importación y exportación	Formatos de los ficheros existentes en los repositorios de datos analizados.	xml, netCDF, odv, hdf, bufr, http, ftp, etc.
Forma de distribución	Sistema utilizado para la difusión de los datos.	Disco óptico, Web, FTP, etc.
Esquema de metadatos	Estándares utilizados en metadatos para la descripción del contenido.	DC, etc.

Sistema de consulta	Tipología y prestaciones de consulta para acceder a los paquetes de datos deseados.	Búsqueda libre, búsqueda por metadatos, sin sistema de búsqueda.
Interoperabilidad	Protocolos que se cumplen para facilitar la interoperabilidad.	ISO, HTTP, FTP OGC, WMS, WFS, CS-W, SWE, OpenSearch, OpenID, Shibboleth, etc.
Identificador	Inclusión de algún identificador persistente de objeto digital.	DOI, ARK, XRI, Handle, LSID, OID, PURL, URI/URN/URL, UUID
Política de acceso	Descripción de la política de consulta de los datos.	Abierto, semiabierto, abierto con registro previo, privado
Derechos de explotación	Especificación sobre el copyright.	Tipo de licencia utilizada
Costes	Existencia, o no, de costos para acceder a los datos.	Gratuito, suscripción, pago por datos, etc.
Estadísticas	Inclusión de estadísticas de uso.	Sí, No

Tabla 1: Indicadores para análisis de los repositorios de datos
Fuente: elaboración propia

La información recopilada se ha obtenido a partir de la consulta externa al repositorio.

Hay que mencionar que en este estudio no se incluyen los datos polares.

Los repositorios analizados han sido los siguientes:

REPOSITORIO	DESCRIPCIÓN
Australian Ocean Data Center Facility (AODC) http://portal.aodn.org.au/aodn/	Creado en 1964, es el NODC de Australia y tiene su sede en Tasmania. Trabaja conjuntamente con seis agencias gubernamentales de datos marinos australianos.
Banco Nacional de Datos Oceanográficos (BNDO) http://www.mar.mil.br/dhn/chm/Oceanografía/bndo.html	Creado en 1994, es el NODC de Brasil, con sede en rio de Janeiro con la supervisión del Centro de Hidrografia da Marinha (CHM)
British Oceanographic Data Centre (BODC) http://www.bodc.ac.uk/	Creado en 1969, es el NODC del Reino Unido, está situado en Liverpool y forma parte del National Environment Research Council (NERC).
Centro Nacional de Datos Oceanográficos do México (CENDO) http://cendo.ens.uabc.mx/	Creado en 2011, es el NODC de México, con sede en la UABC (Ensenada, México).
European Marine Observation and Data Network (EMODnet) http://www.emodnet.eu/	Puesto en marcha en 2009 por la Dirección General de Asuntos Marítimos y Pesca (DG MARE) de la Comisión Europea. Su sede está en Ostende y cuenta con más de 100 participantes.

REPOSITORIO	DESCRIPCIÓN
Geological and Geophysical Data (Geo-Seas) http://www.geo-seas.eu/	Proyecto desarrollado entre 2009 y 2013, el Coordinador del Proyecto es Natural Environment Research Council-British Geological Survey (NERC-BGS), en Nottingham (Reino Unido) y el coordinador técnico es MARIS (La Haya). Se trata de una infraestructura de 26 centros de datos ubicados en 17 países marítimos europeos.
JERICO http://www.jerico-fp7.eu/datatool/	Creado en 2011. Cuenta con 27 participantes y el coordinador es IFREMER .
Integrated Marine Observing System (IMOS) http://imos.org.au/home.html	Creado en 2007, tiene su sede en la Universidad de Tasmania y está formado por diez instituciones del National Research Infrastructure for Australia (NCRIS).
MyOcean www.myocean.eu/	Creada en 2009 y dirigida por Mercator Ocean (IFREMER). Sociedad de organismos públicos y privados compuesta por 59 miembros de 28 países.
National Oceanic and Atmospheric Administration (NOAA) http://www.nodc.noaa.gov	Creado en 1961, es el NODC de los EE.UU., con sede en Silver Spring (Maryland).
Ocean Data Portal (ODP) http://www.oceandataportal.org/	Creado en 2013, a través de la red de NODC del programa International Oceanographic Data and Information Exchange (IODE) (COI UNESCO).
Rolling Deck to Repository (R2R) http://www.rvdata.us/	Creado en 2008, con sede en el University-National Oceanographic Laboratory System (UNOLS) (Rhode Islands).
SeaDataNet http://www.seadatanet.org/Data-Access	Consorcio de 49 instituciones de 35 países creado en 2004. Está coordinado por IFREMER (Brest) y el coordinador técnico es MARIS (La Haya).
Centro Argentino de Datos Oceanográficos (CEADO); Servicio de Hidrografia Naval http://www.hidro.gov.ar/ceado/ceado.asp	Creado en 1974, es el NODC de Argentina, con sede en el SHM (Buenos Aires).
Systèmes d'Informations Scientifiques pour la MER (SISMER) http://www.ifremer.fr/sismer/index_UK.htm	Creado en 1990, es en Francia el NODC (Centro Nacional de Datos Oceanográficos), desarrollado por la unidad de investigación informática y datos marinos de IFREMER.

Tabla 2: Bases de datos analizadas
Fuente: elaboración propia

Finalmente, se ha llevado a cabo la evaluación de los repositorios en base a los indicadores establecidos. Toda la información recopilada se ha obtenido a partir de la consulta externa al repositorio a pesar de que algunos indicadores son difíciles de evaluar mediante este sistema. Los datos recopilados, por tanto, no proceden de cuestionarios enviados a los responsables de los repositorios.

A continuación vamos a presentar la relación de los repositorios analizados así como la valoración general en base a los indicadores establecidos en el Apéndice A.

4.4.3 Resultados

— Ámbito geográfico

Figura 21: distribución geográfica de las bases de datos internacionales
Fuente: elaboración propia (2015)

El panorama internacional demuestra que la gran mayoría de los repositorios oceanográficos están situados en Europa y América del Norte. Estos continentes presentan una infraestructura adecuada para el intercambio de datos entre diferentes plataformas, incluso sirviendo como recolectores de datos procedentes de centros de investigación de América Latina, la cual está representada con dos repositorios, así como Oceanía. Por otro lado, el 60% de los repositorios analizados son de organismos estatales —como pueden ser IFREMER (FR), MARIS (HOL), Natural Environment Research Council-British Geological Survey (NERC-BGS) o NOAA (EE.UU.), por citar algunas— mientras que el 40% restante están situados en organismos internacionales como UNESCO-COI-IODE o la DG Asuntos Marítimos y Pesca de la Comisión Europea (CE).

Precisamente se tiene que destacar el papel de IFREMER, que participa de 2 consorcios y 1 centro nacional de datos, así como el MARIS, que participa

directamente en el desarrollo de 2 consorcios y la participación de Australia en la infraestructura de 2 agencias de gestión de datos oceanográficos.

— Tipo de servicio

Existen 4 repositorios (29% del total) que actúan como proveedores de datos: Rolling Deck to Repository, Centro Argentino de Datos Oceanográficos (CEADO), NOAA y el Banco Nacional de Datos Oceanográficos (BNDO). En el resto de los casos, se trata de agregadores, de servicios que reúnen datos procedentes de diversos repositorios para facilitar una búsqueda unificada y ofreciendo también formatos comunes de indexación de datos, y facilidades para el intercambio de datos.

— Tipo de datos

En general todos los repositorios reúnen datos de todas las disciplinas presentes en la oceanografía: oceanografía física (93,3%), oceanografía química (93,3%), geología marina (100%), biología marina (86,6%). Los datos biológicos no se usan en la infraestructura paneuropea Geo-Seas (de carácter monográfico en la geología), tampoco los emplea la organización militar brasileña responsable de BNDO (centrada en la hidrografía).

— Formato de importación y exportación

Entre todos los repositorios analizados, hay una notable variedad de formatos (40 diferentes) siendo los más importantes NetCDF (73,3%), ASCII (33%), XML (20%), ODV (20%) y CSV (20%). De los alrededor de 300 formatos con los que se manejan datos marinos, en los 14 repositorios estudiados un gran volumen se estandariza alrededor de un único formato (NetCDF). Hay que recordar, sin embargo, que lo mejor cuando hay tantos estándares es que se puede escoger entre ellos (Tanenbaum, 1981, p.168).

— Fuente de datos

Los datos proceden de centros de investigación y, en la mayoría de casos, de consorcios que intercambian datos entre ellos por medio de la adopción de políticas comunes del uso de formatos para agregar y dar acceso a sus archivos.

— Disponibilidad

Sobre los costos para acceder a los datos, los repositorios recolectores disponen de acceso para metadatos que apuntan los paquetes de datos que disponen.[, pero en general es necesario pagar para accederlos por completo.] Por lo normal, se reconoce que los datos marinos son un bien público, estipulándose una política de acceso gratuito con excepciones. La política de copyright del IFREMER distingue entre costes marginales para el acceso con propósito científico y una estimación general de los costes antes de la entrega de los datos cuando el propósito es general. Hay también datos públicos con derecho de privacidad, siendo necesario efectuar registro para visualizarlos.

— Forma de distribución

Como los volúmenes de datos oceanográficos son muy grandes y, por su propia naturaleza, descentralizados, los sistemas que analizamos ofrecen diversos modos de gestionar su distribución. La descarga desde una página o portal web está disponible de forma general en el 93% de los repositorios, y también es posible vía FTP (40%), correo electrónico (20%), CD-ROM (20%) o DVD (20%).

— Sistema de consulta

La consulta sólo es posible a partir de los metadatos que se han definido. La excepción es Ocean Data Portal, donde es posible hacer búsqueda directamente por el numero del registro de un dataset. Así mismo la mayoría no presenta la posibilidad de acceder directamente a ningún archivo, mientras algunas bases de datos utilicen el protocolo CSW (Catalog Service for the Web), como recurso para especificar un patrón de diseño para la definición de interfaces para la publicación y búsqueda de colecciones de información descriptiva (metadatos) sobre datos geoespaciales, servicios y objetos de información relacionada. Las bases

presentan sistemas de búsqueda simples y avanzada, además hay posibilidad de consulta por medio de tesauros, directorios propios, áreas de investigación oceanográfica, tales como batimetría, geologia, etc, excepto el Rolling Deck to Repository, lo cual no dispone interface de búsqueda y solamente informaciones sobre los datos de crucero que reúne. Todavía el Centro Nacional de Datos Oceanográficos de México también mantiene un sistema de búsqueda atípico: dispone resultados de informes generales de los Datasets por asunto y acompañados de informes descriptivos, al envés de disponibilizar los conjuntos de datos. En relación al Centro Argentino de Datos Oceanográficos y el Banco Nacional de Datos Oceanográficos (Brasil), no disponen servicio de búsqueda en línea y las consultas deben ser realizadas por correo electrónico o teléfono. En el análisis de las bases, destacamos el Australian Ocean Data Center Facility, la cual dispone resultados para visualización directamente en Atlas geográfico correspondente a la región de la consulta.

— Tipo de servicio

Con excepción de las bases de datos Rolling Deck to Repository, Centro Argentino de Datos Oceanográficos (CEADO) y el Banco Nacional de Datos Oceanográficos (BNDO), todas los demás son agregadores, reúnen datos de una variedad de bases de datos para busca centralizada. Las principales ventajas de las bases agregadores son los formatos comunes de indexación de datos, el intercambio de datos entre ellas y la busca unificada en única plataforma. En relación a las bases de datos recolectoras, aunque trabajen aisladas, envían los datos para bases de datos internacionales, las cuales se encargan de las adaptaciones necesarias.

— Esquema de metadatos

Es necesario garantizar que los datos sean registrados, mantenidos y preservados de manera adecuada. Uno de los requisitos iniciales es que los conjuntos de datos estén acompañados de informaciones que describan cómo se han obtenido (tiempo o espacio, métodos e instrumentos de recogida), cuál es su ámbito, autoría, propiedad y condiciones de reutilización, control de calidad con el horizonte de un control por pares similar al que funciona en el caso del arbitraje de

los artículos científicos, etc. A este conjunto de descriptores se les denomina metadatos. Así, juntamente con la interoperabilidad tecnológica, la existencia de metadatos adecuados y normalizados es un requisito esencial para el acceso y reutilización de datos científicos (Schaap & Glaves, 2014).

Los esquemas de metadatos no son especificados en todas las bases de datos. Las que disponen su organización están divididas por áreas, en física, química y geofísica, geología y biología, en formatos numérico y textual, agrupados por área geográfica. El análisis de las bases de datos demuestra una padronización sobre el formato de las fechas (YYYYMMDD) y Hora (formato hhmmss), de acuerdo con las recomendaciones de la Intergovernmental Oceanographic Commission (COI) a través de la norma ISO 8601: 2004 Ocean Data Standards

El formato de fichero NetCDF en la mayoría de las bases es utilizado para describir los conjuntos de datasets. El formato NetCDF más comun consta de un cabecero, que contiene toda la información sobre las dimensiones, atributos y nombres de las variables contenidas y la parte de los datos, comprendiendo los datos de las variables de tamaño fijo, y los datos de tamaño variable, conteniendo los datos de las variables que tienen dimensión ilimitada. El modelo más completo (netCDF-4) permite estructuras más complejas de datos con tipos de datos definidos por el usuario, permitiendo almacenarlos de forma jerárquica mediante grupos (que serían como carpetas en un sistema de ficheros).

Los estándares de metadatos centrales para los datos oceanográficos en general son el modelo ISO 19115 (contenido).

En relación al esquema de metadatos para descripción de datasets de investigaciones en la Antártida, es usado el formato Directory Interchange Format (DIF), como es el caso de la National Oceanographic Data Center (NODC-NOAA). Un DIF se compone de ocho campos que detallan información específica sobre los datos y otros campos que amplían y aclaran la información.

— Interoperabilidad

Las colecciones de mapas digitales y los conjuntos de datos con tablas complementarias, figuras e informaciones que ilustran de manera sistemática la costa como para la gestión de las zonas costeras y la planificación, a menudo con herramientas cartográficas y de toma de apoyo, todos los cuales son accesibles a través de Internet (formatos FTP y HTTP) o por solicitación. Un índice basado en la ISO 19115 a los datos de las muestras individuales, núcleos y mediciones geofísicas y una interfaz única para acceder a estos conjuntos de datos en línea.

— Identificador

Únicamente en SISMER, accesible a través del portal para los datos de Ifremer, se estipula que para valorizar al productor del dato, y facilitar la cita, se puede asociar a cualquier conjunto de datos un DOI (Identificador de Objeto Digital). Para ello debe constar una solicitud en el momento del depósito de los datos. En el resto de los repositorios no existe la posibilidad de asignar un identificador.

— Políticas de acceso

Los usuarios de los repositorios son capaces de acceder en modo abierto en un 40% de los casos, sólo encontrando una limitación derivada del registro previo en el 27% de las infraestructuras. Hay datos sujetos a condiciones específicas de acceso y uso, de carácter semiabierto, en un 20% de los repositorios y, finalmente, en un 13% la información no está disponible.

— Costes

Los repositorios analizados proporcionan acceso gratuito a los datos del océano en el 53% de los casos. El acceso al 27% de los recursos tiene un coste, aunque es muy flexible. Se facturan la intervención manual, los cargos para terceros, los productos CD-ROM/DVD, datos para el cliente y de archivo, y los costes marginales por uso científico. El 20% de las plataformas no ofrece información.

— Tipo de licencia

Los centros productores de los repositorios (la universidad, los centros nacionales de datos oceanográficos, los establecimientos públicos de carácter industrial y comercial) ostentan los derechos de propiedad intelectual para crear, compartir y reutilizar los datos, en el 33% de los repositorios analizados. Las propias infraestructuras tienen derechos específicos de copyright sobre el material en igual proporción (33%). La identidad de los detentores del derecho se desconoce en el 27% de los casos. Y en un único caso las condiciones de acceso y uso están reguladas por una licencia de atribución Creative Commons.

— Estadísticas

Systèmes d'Informations Scientifiques pour la MER (SISMER) es la única base de datos que dispone informaciones sobre estadísticas con informes anuales. MyOcean ofrece estadística por área de búsqueda en cualquier momento y SEADATANET dispone el servicio DIVA (Análisis de datos variacional-interpolación), un software para el análisis estadístico y la interpolación. Permite la interpolación espacial de estas observaciones en una rejilla regular. Estos campos reticulados pueden utilizarse en numerosas aplicaciones, incluyendo la verificación de la consistencia de las mediciones (es decir, detección de valores atípicos), inicialización, calibración y validación de los modelos oceánicos (en apoyo de proyectos como MyOcean), los análisis de los cambios y tendencias en estacional, anual e interanual escalas de tiempo y análisis de presupuesto (como el contenido de calor y biomasa total). Estos dos repositorios, coordinados por IFREMER, ofrecen pormenorizadas estadísticas bajo clave de autentificación gratuita. Las bases de datos restantes no disponen informaciones sobre el control estadístico de los datos.

4.4.4 Análisis global de los repositorios

El análisis realizado muestra que los datos oceanográficos se encuentran todavía en etapa de maduración en cuanto a su procesamiento, difusión y reutilización. Se constata un interés incipiente en los repositorios, que son un elemento clave para el almacenamiento y la reutilización de estos datos oceanográficos.

Tal y como se ha visto, la mayoría de los repositorios se encuentran en Europa y América. Aunque buena parte de ellos tiene carácter internacional se constata la ausencia de plataformas en Asia y Oceanía. Una buena parte de estos repositorios son agregadores, que se ocupan de chequear regularmente los proveedores de datos buscando nuevo contenido. De esta forma, al recoger y mezclar datos de múltiples fuentes permiten una visión holística de todo el entorno informativo.

El alto potencial de los datos recogidos en los repositorios analizados está asociado a la equilibrada tipología de sus contenidos. Aunque no siempre estén bien gestionados, ni dispongan de asesoramiento en cuanto a su calidad, los datos oceanográficos responden a las necesidades de las distintas comunidades disciplinares (oceanografía física, química marina, geología marina, biología marina).

Los sistemas de consulta presentan las dificultades de rigidez propias de los esquemas de metadatos, profusamente utilizados por todos los repositorios. Sólo tres de ellos permiten expresar el término de búsqueda como un texto libre. Sin embargo, en dos de ellos (SeaDataNet y el británico BODC) los protocolos de control y asesoramiento de la calidad de los metadatos, orientan la búsqueda hacia una sintaxis libre de error, en conformidad con el esquema de metadatos en uso en la sede web. Sólo el servicio de acceso ODP (Ocean Data Portal), que responde a los estándares para los datos del Océano de la red IODE, permite el empleo en la consulta de una frase o sentencia que describa en lenguaje natural la materia por la que se busca.

La mejora de las sedes web de los repositorios analizados figura entre los esfuerzos para promover la reproducibilidad de los datos oceanográficos en la literatura científica publicada y aumentar su transparencia. El objetivo consiste en conseguir que los datos tengan mejor identidad, sean más abiertos y sencillos de acceder, y sobre todo estén mejor documentados.

Finalmente, si se quiere mejorar la situación de la gestión de los datos oceanográficos es fundamental conseguir la implicación de los investigadores y también mejorar el grado de coordinación entre las instituciones que recogen y difunden estos datos.

En primer lugar, pues, se requiere una mayor aceptación en la comunidad académica, sobre todo por parte de los investigadores, muchos de los cuales no difunden ampliamente los datos de sus investigaciones. Todavía es frecuente que los investigadores no envíen los datos a un centro de recogida incluso si conocen el repositorio apropiado. Es una manifestación del escaso interés para la publicación de datos en abierto o de la reticencia en la publicación de datos en abierto. Sin embargo, las sinergias entre revistas científicas, repositorios de datos y revistas de datos ofrecen importantes posibilidades para descubrir y reutilizar los resultados de investigación. Un ejemplo de ello es el atlas mundial de datos del ecosistema marino (MAREDAT) (Buitenhuis, Vogt, Moriarty, Bednaršek, Doney, Leblanc, et al., 2013) (Quadt, Düsterhus, Höck, Lautenschlager, Hense, Hense et al., 2012).

Por otro lado, la falta de coordinación entre instituciones limita la posibilidad de compartir los datos oceanográficos y no permite afrontar de manera adecuada el reto de la investigación abierta. Es fundamental disponer de un entorno de desarrollo rico y flexible donde verificar en la práctica la efectividad de los enfoques en acceso abierto que ya existe (como la atribución de registros DOI a los datos). La infraestructura para los datos espaciales franceses sobre el entorno marino es un ejemplo de buenas prácticas en este sentido (Satra-Le Bris, Quimbert, Treguer, Louarit, 2013) (Merceur, 2014).

En cualquier caso, no se puede olvidar que la gestión de los datos (recolección, procesamiento adecuado, disponibilidad, posibilidad de reutilización, etc.) constituye el gran desafío para el progreso de la investigación oceanográfica y que los repositorios constituyen un elemento fundamental en este proceso.

5 SITUACIÓN EN BRASIL

En este capítulo analizaremos el estado de la arte respecto de la adquisición, procesamiento y presentación de datos oceanográficos en Brasil. Por medio de un análisis del panorama brasileño y de un estudio de usuarios, serán descritas las diferentes formas de adquisición de datos mediante el uso de instrumentos y equipos oceanográficos y la realización de mediciones en el fondo marino y el subsuelo brasileño. Además, presentaremos una valoración de los desafíos del desarrollo para la integración y la sistematización de los datos oceanográficos.

5.1 Antecedentes

Poco después de la llegada de los portugueses a Brasil, su territorio ya estaba siendo representado en cartas náuticas o mapas de rutas que posibilitaban que los colonizadores volviesen a los lugares deseados de la costa. Por lo tanto, se puede decir que la Oceanografía brasileña comenzó con la cartografia. En 1500 ya aparecía la representación de un tramo de la costa brasileña en el diseño de Juan de la Cosa; dos años después, el país fue representado en el planisferio de Cantino y, en 1508, la ruta de Duarte Pacheco Pereira, aporta información valiosa sobre la costa de Brasil.

Las primeras investigaciones sobre los organismos marinos se iniciaron a la primera mitad del siglo pasado, principalmente a través de los naturalistas europeos que vinieron a estudiar y recoger los animales y las plantas de los museos de sus países de origen.

El primer estudio importante en la costa brasileña se llevó a cabo en 1857, cuando la Marina hizo un levantamiento hidrográfico de la desembocadura del río Mossoró (en Río Grande del Norte) hasta la desembocadura del río São Francisco (Alagoas). Este trabajo fue realizado por el capitán de fragata Antonio Manoel de Oliveira Vital, muerto después durante la Guerra del Paraguay y ahora considerado el patrón de la hidrografía brasileña. La institucionalización de un servicio hidrográfico en el país data de 1876, cuando se creó la Oficina de Carta Marítima, actual Directoría de Hidrografía y Navegación Marina (DHN).

La investigación oceanográfica académica, se inició con los trabajos del investigador francés Wladimir Besnard (1890-1960), invitado por el gobierno del

estado de Sao Paulo para organizar el Instituto Paulista de Oceanografía (creado por Decreto-Ley en diciembre de 1946), siendo entonces la primera institución nacional dedicada a la investigación de los recursos vivos, minerales y energéticos del mar brasileño. En 1950, se publicó por el Instituto recién creado, la primera revista nacional en el área de Oceanografía (Marina de Brasil, 2016).

En 1950, el Instituto Paulista de Oceanografía se incorporó a la Universidad de Sao Paulo (USP), recibiendo el nombre de Instituto Oceanográfico. Dos años más tarde, fue contratado por el Instituto, el islandés Ingvar Emilsson, considerado el primer oceanógrafo físico de la enseñanza superior brasileña. Otro personaje importante que contribuyó al desarrollo de la Oceanografía de Brasil fue el Almirante Paulo Moreira da Silva, que convirtió el navío escuela Almirante Saldanha en el primer navío oceanográfico de Brasil, en 1964. En 1967, llegó a Brasil el buque oceanográfico de la Universidad de São Paulo (USP) Profesor Wladimir Besnard, construido en Noruega.

El 5 de enero de 1983, Brasil atracó por primera vez en la Antártida con el navío oceanográfico Profesor W. Besnard y el buque de apoyo oceanográfico Baron Teflé, primer navío polar brasileño, perteneciente a la marina. El objetivo principal de este primer viaje fue iniciar el cumplimiento de las condiciones estipuladas por el Tratado Antártico (1959), y facilitar que Brasil fuera admitido a la posición de membro activo de la comunidad Antártica, con derecho a voz y voto. La estación brasileña en la Antártida, fundada en febrero de 1984, se denominó Comandante Ferraz, en honor a la persona, que jugó un papel importante en el desarrollo de PROANTAR (Programa Antártico Brasileño). Se instaló en la bahía del Almirantazgo, al lado de la isla Rey Jorge, cerca de la Península Antártica. La Comisión Interministerial para los Recursos del Mar (CIRM) creó el subcomité PROANTAR con representantes de diversas entidades, con el encargo de preparar el programa, sobre la base de proyectos recibidos de diversas instituciones.

5.2 Formación universitaria

En relación a la formación universitaria, existen tres niveles de enseñanza en el escenario brasileño que ofrecen capacitación para investigación científica en Oceanografía.

a) Primer ciclo

Existen pocos grados universitarios en oceanografía: la Universidad Federal del Rio Grande (FURG), que comenzó en 1971 (el año del ingreso de la primera clase), fue el precursor de esta área en el país, lo que provocó el primer ciclo de la creación de estos cursos, que se extendió a partir de los años 90. En este período se crearon los cursos de la Universidad del Estado de Río de Janeiro - UERJ, en 1977 con mayor énfasis en la oceanografía biológica; también con un mayor énfasis en esta área, además algun soporte en la Oceanografía Física y en la Universidad del Vale do Itajaí (UNIVALI), en Santa Catarina, en 1992, destacando también la oceanografía biológica. Más recientemente se han creado cursos en el Centro de la Universidad Monte Serrat (en la ciudad de Santos), en la Universidad Federal de Espírito Santo y en la Universidad Federal del Pará. En 1994 fue creada por el Instituto de Física y el Instituto Oceanográfico de la Universidad de Sao Paulo, una licenciatura en Física con especialización en oceanografía física. Hay cursos más allá de los ya mencionados, otros que tienen grados con asignaturas relacionadas con el medio marino como las Ciencias Biológicas, de la Universidad de Sao Paulo, o de la Universidad de Santa Cecilia (Santos).

La creación de cursos de grado fue decisiva para el desarrollo de la oceanografía en Brasil y refuerzan en la convicción de que el mar era una fuente inagotable de recursos, la pesca, en particular, por lo que existía la necesidad de la formación de un profesional capaz de contribuir a la explotación de estas riquezas.

b) Segundo ciclo

Se inició gracias a la entrada en vigor de la Ley Nº 9394 (Ley de Directrices y Bases de la Educación - LDB) del 20.12.1996, que aseguró las instituciones reconocidas como universidades y centros de la autonomía universitaria para crear cursos de posgrado. Hasta entonces, todas las instituciones tenían que

solicitar la autorización previa del Ministerio de Educación - MEC. En un momento en que la preocupación por los problemas medioambientales comenzó a ganar cada vez más espacio en la sociedad, a la que contribuyó en gran medida a la Conferencia de las Naciones Unidas sobre el Medio Ambiente y el Desarrollo - CNUMAD (ECO 92, o RIO 92), la creación de nuevos cursos, que se centraron principalmente en la conservación y el uso sostenible de los recursos hídricos, se produjo como un proceso natural. De esta manera se crearon los cursos de Oceanografía en el Centro de la Universidad Monte Serrat - UNIMONTE (São Paulo) en 1998, la Universidad Federal de Espírito Santo - UFES y la Universidad Federal de Pará - UFPA en 2000, la Universidad de Sao Paulo - USP en 2002 y también en la Universidad Federal de Bahia - UFBA y la Universidad Federal de Paraná - UFPR en el año 2004.

c) Tercer ciclo

Existen cinco instituciones nacionales que llevan a cabo cursos en diversas áreas de Oceanografía: Universidad Federal del Rio Grande, Instituto Oceanográfico de la Universidad de Sao Paulo, de la Universidad Federal de Rio Grande do Sul, de la Universidad Pontificia de Río de Janeiro y la Universidad Federal del Pernambuco.

Aún en curso, el tercer ciclo de la creación de cursos de Oceanografía fue provocada por la decisión del gobierno para implementar el Programa de Apoyo a la Reestructuración y Expansión de las Universidades Federales (REUNI). En este contexto, algunas instituciones que tenian grupos de investigación y/o programas de postgrado en el área de Ciencias del Mar, trataron de aprovechar las condiciones favorables y propusieron la creación de nuevos cursos en Oceanografía. Así, en 2008, se crearon los cursos de la Universidad Federal de Santa Catarina - UFSC y la Universidad Federal de Ceará - UFC y, en 2009, en la Universidad Federal de Pernambuco - UFPE.

Aunque no estén relacionados con el contexto REUNI, otro curso de Oceanografía entró en funcionamiento a partir de 2010 en la Universidad Federal de Maranhão - UFMA como resultado del cambio de denominación del curso Ciencias Acuaticas, desplegado en esa institución 2002. Con la entrada en

actividad de los nuevos cursos son 11 los estados costeros brasileños que tienen al menos un curso de Oceanografía.

5.3 La investigación en Oceanografía

Brasil tiene un papel muy relevante en la investigación oceanográfica del Atlántico Sur. Con una costa de 7.408 kilómetros, aproximadamente el 40% de las regiones costeras del Atlántico Sur pertenecen a Brasil. El Atlántico Sur es de gran importancia debido al flujo de calor significativo entre la Antártida y la región ecuatorial, jugado papel crucial en el sistema climático del planeta. En Brasil, por ejemplo, hay indicios de que el fenómeno cíclico de las sequías del noreste se asocia con anomalías de la temperatura superficial de las aguas de la región ecuatorial del Océano Atlántico. Además, hay cambios atmosféricos en el hemisferio sur aún poco conocidos, que influyen en el clima brasileño, tales como El Niño se produce en la costa oeste de América del Sur.

Actualmente en Brasil existen instituciones que trabajan en el campo de la oceanografía en diferentes niveles de participación, desde el totalmente dedicado a las ciencias del mar a los que tienen pequeños grupos que desarrollan actividades en el campo de la Oceanografía. Por otra parte, hay de 419 investigadores que se dedican a actividades oceanográficas, de acuerdo con el Consejo Nacional de Desarrollo Científico y Tecnológico (CNPq):

	Doctores en Oceanografía (Investigacion y enseñanza):	Mestres en Oceanografía (Investigacion y enseñanza):	Doctores en Oceanografía (Administrativas, técnicas y otras):	Mestres en Oceanografía (Administrativas, técnicas y otras):
Sudeste	6	3	11	32
Norte	0	0	0	0
Nordeste	5	0	0	3
Centro Oeste	6	2	4	25
Sur	62	24	17	71

Tabla 15: Investigadores en la Oceanografía brasileña
Fuente: Extracción de datos de la base del Currículo Lattes (Painel Lattes)

En lo que respecta a la producción científica, Brasil presenta notorios avances en el campo de las ciencias marinas. El SCImago Journal & Country Rank, un portal que incluye indicadores científicos de las revistas a partir de la información contenida en la base de datos Scopus, y que apunta indicadores pueden ser utilizados para evaluar y analizar los dominios científicos, muestra que Brasil ocupa el 18 lugar en el ranking mundial de publicaciones científicas en Oceanografía. Con una relevante presencia en el escenario internacional, es lider en producción científica oceanográfica en Latinoamérica.

	Country	Documents	Citable documents	Citations	Self-Citations	Citations per Document	H index
1	United States	60.855	59.492	1.388.089	786.179	23,81	255
2	United Kingdom	17.138	16.613	337.352	102.441	21,40	153
3	Germany	13.724	13.494	288.142	91.245	22,82	150
4	France	12.879	12.647	282.774	89.174	23,78	155
5	China	11.588	11.460	105.094	62.963	13,58	103
6	Japan	10.876	10.719	160.565	54.363	16,12	118
7	Canada	9.918	9.700	202.358	54.537	22,06	136
8	Australia	8.883	8.695	152.153	53.254	20,96	121
9	Russian Federation	8.423	8.361	62.946	17.941	7,51	84
10	Italy	6.761	6.597	113.331	41.380	19,03	106
11	Spain	6.660	6.521	105.835	41.348	19,95	98
12	Netherlands	5.292	5.195	109.672	23.375	22,75	108
13	Norway	4.569	4.455	86.598	25.016	22,33	97
14	India	4.112	4.058	31.804	16.031	10,04	62
15	Taiwan	3.228	3.178	31.416	12.369	11,96	65
16	Sweden	2.915	2.856	57.290	12.326	21,83	90
17	South Korea	2.831	2.794	25.561	7.595	12,92	60
18	Brazil	2.658	2.576	33.115	11.972	18,79	74

Tabla 17: Ranking producción científica internacional en Oceanografía
Fuente: SCImago Journal & Country Rank (2016)

5.4 Estudios polares

Los estudios polares constituyen una especialización relevante dentro de la oceanografía. En cuanto a la investigación en la Antártida, Brasil ha demostrado que es un miembro activo del Tratado Antártico, y ha logrado proyectos científicos de gran relevancia. El Programa Antártico Brasileño (Proantar) ha

producido investigación de alta calidad en más de 30 años de acción[28], permitiendo que la ciencia polar brasileña entrara en el escenario científico internacional de forma muy competente. Sin embargo, para la investigación en el entorno polar, en el que las dificultades pueden tener consecuencias costosas para las instituciones de fomento de la ciencia, la gestión adecuada de los datos que representan las actividades y la investigación científica debe ser una ventaja científica brasileña a ser alcanzada.

El Proantar ha impulsado la investigación destinada a aumentar el conocimiento sobre los fenómenos que se producen en el continente y el Océano Austral en todos sus aspectos y su influencia en el Brasil (CNPq, 2011), cooperando activamente con el Scientific Committee on Antarctic Research (SCAR)[29], una organización internacional, interdisciplinaria y no gubernamental que desarrolla y gestiona la investigación científica de alta calidad en la región antártica (CGEE, 2006; SCAR, 2011). Este comité es parte del International Council for Science (ICSU) y fue creado a raíz de la conclusión del Año Geofísico Internacional en 1957-58 (MACHADO; BRITO, 2006; SCAR, 2011).

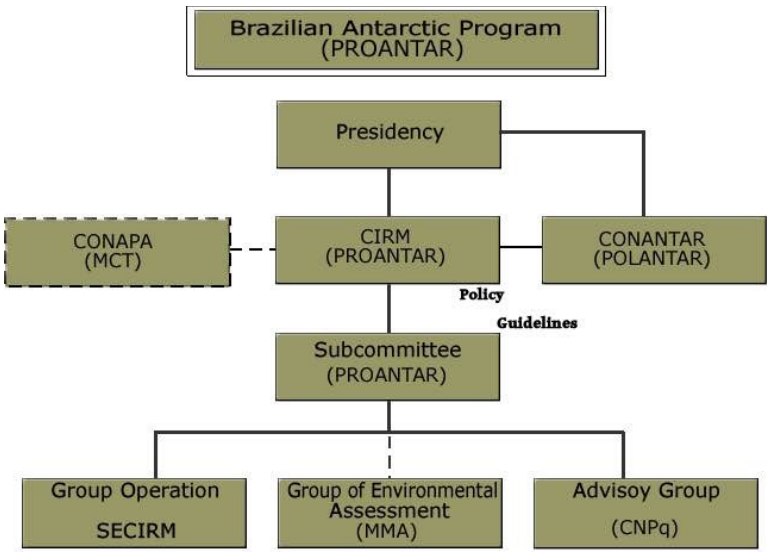

Figura 22: Sectores responsables por la gestión de datos polares en Brasil
Fuente: PROANTAR (2014)

28 De acuerdo con el Programa Antártico Brasileño, Brasil a la Antártida fue la primera vez en el verano de 1982.

29 http://www.scar.org/

En Brasil, las pocas fuentes de información sobre el área de la ciencia polar se encuentran dispersas. Paradójicamente la bibliografía es muy extensa, lo que permite una mirada panorámica en estudios antárticos. No hay duda de que el Proantar fomentó la investigación de alta calidad, sin embargo, no aparece en los sitios web de las instituciones responsables de este programa, la información relativa a la producción científica generada por los proyectos de la Antártida (González, 2010). En algunos casos, contenía sólo la producción de proyectos específicos en los sitios web de los grupos de investigación que participan en expediciones antárticas. Vale la pena mencionar que la producción científica brasileña, como resultado de las actividades en el Polo Sur, tiende a ser entendida como una forma de medir los resultados obtenidos por Proantar a lo largo de los 30 años de actividad del programa.

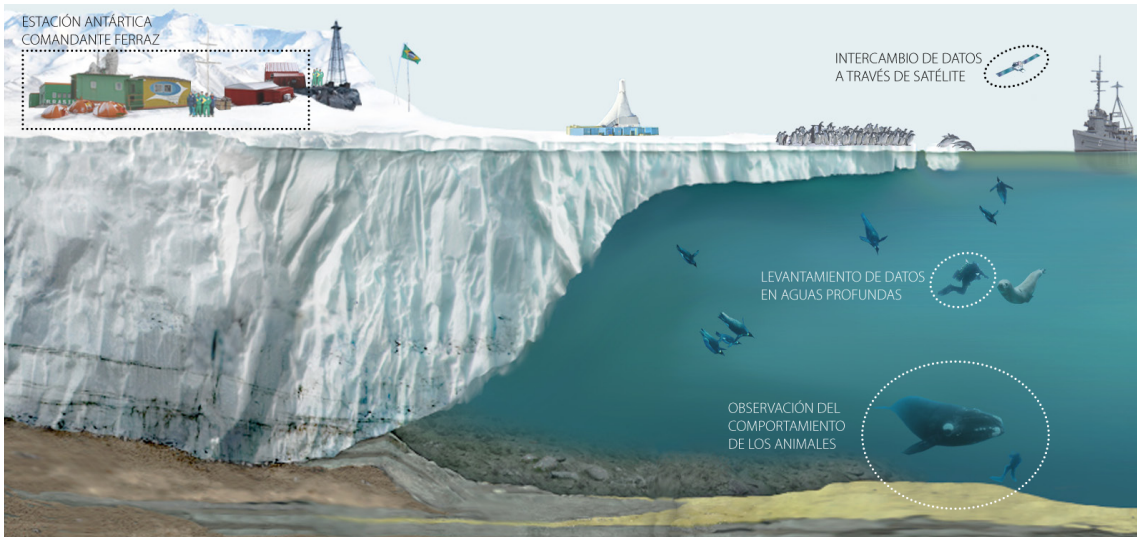

Figura 23: Principales fuentes de datos de la investigación brasileña en la Antártida
Fuente: elaboración propia (2014)

En 2006 el CGEE celebró a petición del CNPq, una evaluación de la investigación científica Proantar (CGEE, 2006). El informe es una de las pocas fuentes de información general sobre el programa de investigación, lo que permite, dentro de las posibilidades, una visión temporal y temática de lo que ha sido realizado y producido desde 1982 hasta 2005. El informe indicó que, a través del CGEE, fue desarrollado internamente un sistema de información para el conocimiento más amplio del Proantar (CGEE, 2006).

Para analizar los datos sobre la producción científica de la investigación llevada a cabo en el Polo Sur es necesario describir las características de la productividad,

tanto nacionales como internacionales, en grandes áreas de los estudios polares. Por lo tanto, con el objetivo de medir la relevancia de las humanidades, las ciencias sociales, la información y la salud en el contexto de la ciencia polar y el papel de los sistemas de almacenamiento de datos, llevamos a cabo un análisis bibliométrico para investigar las tendencias y características de investigación en la Antártida.

5.4.1 Método y fuente de los datos

Para medir la producción científica brasileña fue adoptada la evaluación desarrollada por el CGEE sobre las investigaciones relacionada con el Proantar. En esta evaluación se incluyen publicaciones en revistas, presentaciones en conferencias y publicaciones de libros y capítulos de libros. Por lo tanto, el informe del CGEE fue utilizado para la comparación con los datos internacionales, teniendo en cuenta la ausencia de un NADC brasileño. Vale destacar que el informe CGEE (2006) agrupó los datos de acuerdo con la clasificación adoptada por el SCAR. Además, con el fin de complementar y actualizar la información contenida en el informe CGEE también se consultó libros y las direcciones electrónicas de las instituciones que participan en la investigación polar brasileña. Los datos recogidos fueron sometidos a un análisis descriptivo de la frecuencia absoluta y relativa con la ayuda de los programas SPSS19 y Microsoft Excel 2014.

5.4.2 Resultados

En el caso de Brasil, según los datos recogidos en la evaluación del Programa Antártico Brasileño realizado por CGEE (2006), además de presentar variaciones en el número de proyectos y publicaciones que acompañan las inversiones en investigación y el desarrollo de eventos temáticos, cuenta también con una producción de 1300 publicaciones en 23 años y muestra una clara tendencia creciente de la producción científica polar lo largo de los años. Sin duda, los científicos brasileños han llevado a cabo investigación antártica de alta calidad y disponen de reconocimiento a nivel internacional. Hay que destacar, además, la

inversión del equivalente de 25 millones de reales a 540 proyectos financiados por el Proantar durante este período (CGEE, 2006).

Los medios de difusión científica más utilizados fueron la presentación en eventos nacionales, seguido de presentaciones en conferencias internacionales. En relación a las publicaciones de artículos en revistas nacionales e internacionales están en el orden del 8% y 9%, respectivamente. Como puede verse en el Gráfico 3, los estudios brasileños publicados en revistas científicas internacionales fueron significativamente mayores en el área de ciencias de la vida, con un 57% de las publicaciones, ya que el área de ciencias de la tierra representaron el 30%, seguido por el 13% las ciencias físicas. Además, la producción científica total por área del conocimiento se observa que ciencias de la vida volvió a ser el más productivo (58%), seguido de ciencias de la tierra (24%) y las ciencias físicas (15%). El item Otro representa sólo el 3% de la producción total. El número de proyectos y los recursos asignados para la 65 zona tienen porcentajes similares entre sí, pero significativamente diferente de la producción científica total.

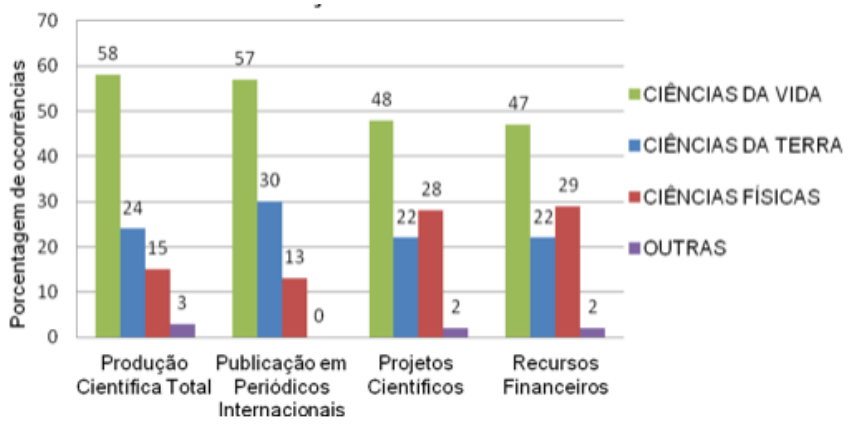

Gráfico 3: Producción Antártica Brasileña. Porcentaje brasileña de producción científica, publicaciones en revistas internacionales, proyectos y recursos financieros divididos según la clasificación del SCAR.
Fuente: elaboración propia (2015).

No hay evaluación del CGEE que haga mención a las ciencias humanas, las ciencias sociales y de salud, medicina o la psicología directamente. Sin embargo, la categoría de ciencias de la vida, como vinculada a la cicatriz, incluye, por ejemplo, parasitología, fisiología, morfología y el comportamiento (SCAR, 2011b). Por lo tanto, estos temas son objeto de estudio del grupo de investigación en la

biología humana y la medicina SCAR (2011b). Sin embargo, los libros publicados tanto en celebración de los 25 años de actividad del Proantar, como las destinadas a hacerlo más accesible al público en general a la producción científica brasileña, no revelan ninguna información sobre las investigaciones en estas áreas llevadas a cabo en el entorno polar.

5.5 Gestión de los datos

En el contexto científico actual, los datos producidos por el aumento de las actividades de investigación oceanográfica en Brasil ganan cada vez más importancia y visibilidad. Por esta razón, el tema de la gestión de datos de investigación se encuentra en el centro de las preocupaciones de la comunidad oceanográfica brasileña. El Banco Nacional de Datos Oceanográficos (BNDO), mantenido y supervisado por la Marina de Brasil, es responsable de la gestión de los datos científicos de los institutos de investigación y universidades del país.

El BNDO colabora con bases de datos que recogen datos oceanográficos de varios países, como la *International Oceanographic Data and Information Exchange Program* (IODE), la *Intergovernmental* Oceanographic Commission (COI), que pertenece a UNESCO y la *Ocean Data and Information Network for Caribbean and South America* (ODINCARSA), además de presentar los datos, metadatos y un resumen de los informes para centros oceanográficos internacionales, incluyendo la *World Data Center for Oceanography* (WDCO) y la *World Meteorological Organization* (WMO), comparte datos meteorológicos provenientes de los formularios utilizados para informar de las observaciones meteorológicas, tales como la forma *Surface Synoptic Observations* (SYNOP)[30], y la *International Hydrographic Organization* (IHO).

El Banco Nacional de Datos Oceanográficos (BNDO), ha establecido las siguientes áreas prioritarias para el archivo:

Tipos de datos	Contenido

[30] Este es un código numérico utilizado para notificar las observaciones meteorológicas hechas por las estaciones meteorológicas de superficie y automáticas.

Física y química oceanográfica	Temperatura, salinidad, oxígeno disuelto y otros equipos, como derivado de CTD, XBT, MBT y otros
Geológicos oriundos de muestras de fundo	Muestras geológicas procedentes de fondo
Medidor de marea	Nivel del mar Heights, constantes armónicas de marea y otra
Correntométricos	Dirección, perfiles de profundidad de la intensidad y de la superficie
Tiempo	**T**emperatura del aire seco y húmedo, temperatura del mar, nubosidad, humedad y otras
GEBCO y batimétricos	**I**ndividual y multihaz

Tabla 18: Tipos de datos oceanográficos en Brasil
Fuente: Banco Nacional de Datos Oceanográficos (BNDO)

El GOOS-BRASIL, componente brasileña de *Global Ocean Observing System*, es un sistema nacional de observación océanica para la recogida de datos, control de calidad, distribución operacional de datos oceanográficos y vigilancia climatológica y oceanográfica en el Atlántico Sur y tropical. Desarrolla actividades de monitoreo del nivel del mar para apoyar la investigación en ciencias del medio ambiente destinadas a la mejora de la planificación económica y social y se divide en cinco módulos:

Módulo I: seguimiento, evaluación, predicción del clima	Hace observaciones del cambio climático en los océanos y monitoreo del nivel del mar con el uso de boyas a la deriva.
Módulo II: recursos marinos vivos	Este módulo está previsto para el seguimiento y análisis de los recursos vivos en el mar desde la perspectiva de explotación sostenible.
Módulo III: océanos saludables	Las actividades en este módulo están relacionados con la evaluación del impacto de las actividades humanas, especialmente las relacionadas con la contaminación marina.
Módulo IV: Gestión costera	Utiliza diferentes tecnologías para el control de las aguas costeras en contraste con los utilizados en las aguas oceánicas.
Módulo V: sistemas meteorológicos y oceanográficos	Este módulo tiene como objetivo desarrollar las funciones de comunicación, transmisión de datos, la modelización y la difusión de productos para permitir la generación de servicios.

Tabla 19: Módulos del GOOS Brasil
Fuente: Elaboración propria (2016)

Actualmente diversos sectores brasileños del gobierno tratan de encontrar soluciones para problemas relacionados con la catalogación, procesamiento y explotación de datos marinos. De manera conjunta buscan estructurar datos e información inter y multidisciplinarios generados por diferentes entidades que estudian el medio ambiente marino. Estas son las discusiones para encontrar un modelo común para el intercambio de datos entre los centros de estudios oceanográficos que culminan en una estructura nacional para archivar y compartir datos.

Las siguientes instituciones están involucradas con el GOOS-Brasil: Secretaria da Comissão Interministerial de Recursos do Mar (SECIRM); Diretoria de Hidrografia e Navegação (DHN); Centro de Hidrografia da Marinha (CHM); Banco Nacional de Dados Oceanográficos (BNDO); Ministério da Ciência, Tecnologia e Inovação (MCTI); Instituto Nacional de Pesquisa Espaciais (INPE); Centro de Previsão de Tempo e Estudos Climáticos (CPTEC); Universidade Federal do Rio Grande (FURG); Instituto Oceanográfico da Universidade de São Paulo (IOUSP)

Instituto Nacional de Meteorologia (INMET); Ministério do Meio Ambiente (MMA)

Instituto de Estudos do Mar Almirante Paulo Moreira (IEAPM); Universidade Federal da Bahia (UFBA)

5.6 Responsables políticos

Dada la reconocida importancia de los océanos en los procesos vitales del medio ambiente, su estudio en los distintos puntos de vista, biológicos y físicos, ha adquirido una especial relevancia en el contexto de las estrategias de investigación en Brasil. Desde hace varios años en Brasil existen órganos que fueron creados para el monitoreo del mar, los cuales han definido los principales retos para el futuro en el campo de las ciencias y tecnologías marinas, así como el papel fundamental que deben desempeñar en el contexto internacional.

En Brasil hay órganos y entidades del gobierno central, de los estados federados y del Distrito Federal, así como fundaciones instituidas por el Poder Público, responsables por la preservación y evolución de la cualidad del registro de los

datos oceanográficos. En relación a las universidades, los repositorios institucionales no mantienen archivos de los datos, aunque algunas dispongan los metadatos con informaciones básicas.

El Ministerio del Medio Ambiente (MMA) es el órgano federal que tiene la atribución de planear, coordinar, supervisionar y controlar la política nacional y las directrices gubernamentales fijadas para el medio ambiente. El MMA está estructurado en Secretarías y Departamentos que tratan con los múltiples temas relacionados con el ambiente y los recursos naturales, entre ellos .

Hay órganos ejecutores que son el IBAMA (Instituto Brasileiro do Meio Ambiente e dos Recursos Naturais Renováveis) y el ICMBio (Instituto Chico Mendes de Conservación da Biodiversidad), con la finalidad de ejecutar y hacer ejecutar, como órganos federales, las políticas y directrices gubernamentales fijadas para el medio ambiente.

En los estados, Distrito Federal y en los municipios de la Federación Brasileña hay órganos similares al IBAMA y el ICMBio, que son Órganos Seccionales y Órganos Locales, respectivamente en los estados y municipios, con la atribución, en las respectivas áreas de jurisdicción política, de ejecución de programas, proyectos y el control y fiscalización de actividades capaces de provocar la degradación ambiental.

También hay Consejos Estaduales y Municipales de Medio Ambiente con atribuciones, en general de, entre otras más, promover la preservación; mejorar la recuperación de la calidad ambiental; coordinar y integrar las actividades ligadas a la defesa del medio ambiente; elaborar normas complementares de protección al medio ambiente; estimular la participación de la comunidad no proceso de preservación, mejoría y recuperación de la cualidad ambiental.

5.7 Estudio de usuarios

Para conocer las prácticas de los investigadores brasileños que trabajan con datos de investigación oceanográficos se diseño un cuestionario compuesto por 15 preguntas. Uno de los aspectos investigados es conocer cuáles son las prácticas actuales de gestión de los datos utilizados por los investigadores y

identificar los hábitos y experiencias actuales utilizadas para preservar y compartir los datos que generan.

5.7.1 Introducción

El presente análisis es el resultado de una encuesta y entrevistas realizadas a un conjunto de investigadores brasileños que trabajan directa o indirectamente con datos oceanográficos y dan una reseña de las tecnologías que han sido utilizadas y aplicadas en centros de investigación de Brasil. Además, muestra la capacidad de los investigadores de recoger, evaluar, visualizar, descargar y analizar datos en las áreas en que actúan al reducir la complejidad inherente a los datos oceanográficos.

Con base en las respuestas obtenidas, además del entendimiento de la revisión de la literatura presentada, mostraremos las principales características tecnológicas, políticas de desarrollo y los formatos de soporte que contemplan las necesidades para el avance de la gestión de datos. En este apartado se incluyen tanto aspectos relacionados con la interfaz de las entrevistas como elementos que se incluyen con los resultados de la análisis de las infraestructuras que forman parte de esta investigación.

No existen muchos estudios de estas características, aunque sea posible encontrar trabajos similares que identifican los hábitos y experiencias de los investigadores para compartir los datos de investigación. El artículo de Aleixandre et. al (2014) es uno de los pocos trabajos que presenta un análisis de los investigadores españoles que actúan en ciencias de la salud en relación con la gestión y el intercambio de los datos brutos de investigación. El resultado de la encuesta aplicada ha permitido conocer las actitudes, motivaciones y amenazas que perciben los investigadores los investigadores españoles en ciencias de salud sobre la gestión y la preservación de los datos brutos de sus investigaciones.

Un estudo similar también es presentado por Tenopir (2011) et. al. Los autores identificaran identificaran las prácticas, barreras y facilitadores de intercambio de datos de intercambio de datos entre investigadores. Como resultado del análisis, identificaran que los científicos no depositan sus datos en forma

electrónica para los demás investigadores, por diversas razones, incluyendo la falta de tiempo y la falta de financiación.

Siguiendo esta misma linea, el *Report on integration of data and publications* (Reilly, 2011), procedente del proyecto ODE (Opportunities Data Exchange) sobre la integración de datos primarios y publicaciones, demuestra la importancia de esta cuestión. Dicho informe menciona los repositorios y las editoriales como las vías de almacenamiento preferidas por los investigadores y pone de manifiesto el deseo de reutilizar datos ajenos y una cierta reticencia a compartir los propios, aduciendo problemas legales. Se señalan las rutas verde y dorada. La primera es el depósito de datasets en repositorios o bancos específicos por disciplinas, mientras que la via dorada consiste en almacenarlos en plataformas editoriales junto a la publicación. Se apunta que relacionar los datos con las publicaciones puede aportar dos ventajas añadidas: contextualizar e interpretar los datos y proporcionar valor tanto a los investigadores que los comparten como a las propias publicaciones. Para las editoriales, los principales problemas detectados son la validación y la preservación, al estar bajo su responsabilidad.

En común, el estudio de Aleixandre et. al (2014) y Tenopir (2011) et. y el informe presentado por Reilly (2011) al identificaran que las barreras para el intercambio de datos y preservación están profundamente arraigadas en las prácticas y la cultura del proceso de investigación, así como los propios investigadores. Un proyecto de investigación similar se llevó a cabo en la Universidad de Colorado. Lage, y Maness (2011) realizaron extensas entrevistas con investigadores y desarrollaron ocho perfiles que representan una agregación de los investigadores de la facultad y de los estudiantes de postgrado que entrevistaron. Estos perfiles revelan que las necesidades, las prácticas y la comprensión de la investigación y de los datos varían ampliamente entre las diferentes disciplinas. Y aunque cada institución, departamento e investigador exhiban cualidades únicas, estos perfiles pueden ser útiles para entender los problemas que surgen cuando se trabaja con los investigadores. Los perfiles diseñados por el equipo de la Universidad de Colorado describen el nivel de interés, la cantidad de apoyo que los investigadores sienten que reciben, los

problemas de almacenamiento de datos a los que se enfrentan y la privacidad que requieren sus datos. Un factor que el equipo observó es la falta de apoyo para el almacenamiento y preservación de los datos, una predisposición positiva hacia el movimiento de acceso abierto y la falta de apoyo a la gestión de los datos durante el proceso de investigación. El equipo de Colorado también señaló que los investigadores de las Ciencias de la Tierra parecían más abiertos a la participación de la biblioteca y a la apertura de datos. Los que trabajan en campos muy competitivos, como el de las ciencias exactas, parecían menos receptivos a la participación de la biblioteca en el proceso de gestión de datos.

Aunque Lage (2011) y sus colegas descubrieron una amplia gama de actitudes hacia el intercambio y la curaduría de datos, hubo algunos puntos en común entre muchos de los investigadores que hablaron. La mayoría de ellos no identificaron sus datos como públicos, aunque eso no significa necesariamente que no estuvieran abiertos a compartirlos con otros interesados, ya que los investigadores, a menudo, quieren mantener un cierto nivel de control sobre los datos que comparten.

El estudio de la Universidad de Colorado también demuestra que los investigadores están de acuerdo con los procedimientos o servicios departamentales para el almacenamiento de datos, que la mayoría de los investigadores tienen algunos subconjuntos de datos científicos que no están siendo mantenidos o conservados con un plan en marcha en la actualidad y que perciben las tareas de gestión de datos como distracciones de sus proyectos de investigación. Estas actitudes revelan que es importante que la gestión de datos sea lo menos complicada posible para los investigadores y que la planificación debe tenerse en cuenta antes de empezar la recogida de los datos.

5.7.2 Método

El cuestionario estuvo disponible *online* con el uso de la herramienta Google Docs durante los meses de julio y agosto de 2014 y tenía 15 preguntas con el propósito de identificar el tipo de datos utilizados, los hábitos y las necesidades de uso de este tipo de información y por el posible interés en la reutilización de los datos. Fue realizada una prueba piloto con dos investigadores anteriormente

al envío del cuestionario para la totalidad del universo de la muestra. En anexo se puede encontrar la encuesta completa.

El estudio tiene un carácter cualitativo y se centra en una muestra no representativa de investigadores oceanográficos que actúan en las principales áreas que los estudios en Oceanografía en nivel global, siendo[31]: la física, la biología, la química y la geología. De hecho, seleccionamos expertos de los principales centros del país, con la participación del Norte, Nordeste, Sudeste y Sur de Brasil, según la siguiente tabla:

Norte	Nordeste	Sudeste	Sur
Universidade Federal do Pará (UFPA)	Universidade Federal do Ceará (UFC)	Universidade Federal do Espírito Santo (UFES)	Universidade Federal do Rio Grande (FURG)
	Universidade Federal da Bahia (UFBA)	Universidade Federal do Rio de Janeiro (UFRJ)	Universidade Federal do Estado do Rio Grande do Sul (UFRGS)
	Universidade Federal do Maranhão (UFMA)	Universidade Estadual do Rio de Janeiro (UERJ)	Universidade Federal de Santa Catarina (UFSC)
	Universidade Federal do Pernambuco (UFPE)	Universidade de So Paulo (USP)	Universidade Federal do Rio Grande (FURG)
			Universidade do Vale do Itajai (UNIVALI)

Tabla 3: Universidades que hacen parte de la investigación
Fuente: elaboración propia

El levantamiento y la delineación de los investigadores principales se llevó a cabo con la premisa principal de disponer de una productividad científica de alto nivel registrado en la Plataforma Lattes[32].

Después de terminar el plazo de envío de las respuestas, seleccionamos tres investigadores que poseen productividad científica de alto nivel registrado en la Plataforma Lattes.[33] A ellos preguntamos, vía Skype, la visión critica que tenían

31 http://www.cnpq.br/documents/10157/186158/TabeladeAreasdoConhecimento.pdf

32 http://lattes.cnpq.br/

33 http://lattes.cnpq.br/

sobre cuestiones relacionadas con la infraestructura de datos oceanográficos en el escenario brasileño en relación a:

a) Infraestructura para cargar y disponibilizar los datos;

b) Como Brasil relaciona las investigaciones internas del país con las internacionales.

Las entrevistas se llevaran a cabo entre octubre de 2014 y mayo de 2015, a partir de cuestionarios semiestructurados, e interpretados según la técnica de análisis de contenido. Para esto, se procedió a la lectura de las declaraciones, seguido de la delimitación de fragmentos de las mismas con significado pertinente al objetivo del trabajo. Las unidades de registro fueron agrupadas en categorías, conforme la proximidad del significado que contenían.

El análisis de los investigadores que fueron entrevistados están identificados como “Entrevistado 1”, "Entrevistado 2” y “Entrevistado 3” para mantener el anonimato de los mismos, a pedido de ellos.

Los comentarios de los tres entrevistados se han intercalado en los resultados de la encuesta, para seguir de manera más ordenada un único hilo argumental de discusión.

El discurso de introducción para las entrevistas si encuentran en letra normal y las respuestas de los entrevistados si encuentran en letra cursiva.

5.7.3 Resultados

Del total de las 16 universidades seleccionadas, obtuvimos 42 respuestas: UFSCAR, USP (4), UFRJ (3), UFPR (3), UFBA (2), UFRGS, UFC, UFPE (2), UFSC, UFMA (3), FURG (21). Adicionalmente, contestaron dos institutos del gobierno brasileño que tienen sectores que trabajan con datos oceanográficos: el Instituto Chico Mendes de Conservação da Biodiversidad (ICMbio) y el Instituto Nacional de Pesquisas Espaciais (INPE), sumando el total de 44 respuestas para la encuesta.

5.7.3.1 Producción de datos de investigación

De acuerdo con la encuesta realizada, los investigadores que respondieron la encuesta trabajan con datos primarios en las siguientes áreas: Oceanografía Biológica (43,2%), seguido de la Oceanografía Física (29,5%), encontrándose la Oceanografía Geológica (27,3%), en seguida la Oceanografía Química (9,1%) y otras áreas que no fueron especificadas (4,5%).

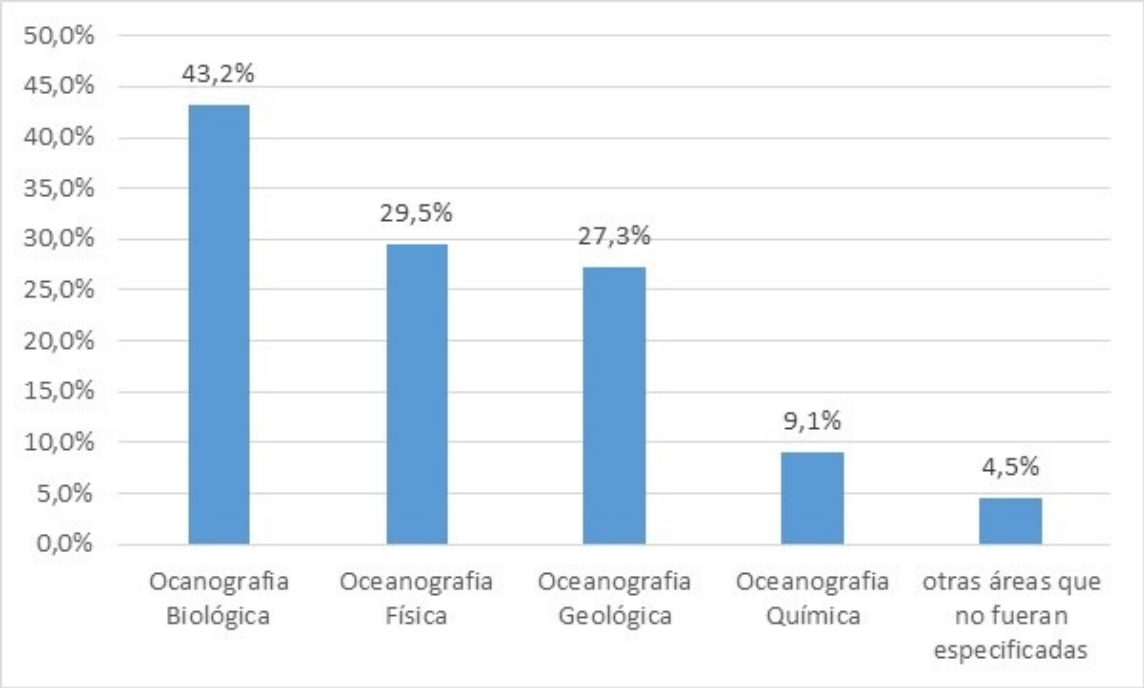

Gráfico 1: Producción de los datos de investigación
Fuente: elaboración propia (2016)

Los diferentes datos producidos por la comunidad de investigadores son muy variadas, ya que incluyen mediciones realizadas por satélites de sensores remotos, observaciones de instrumentos *in situ* y análisis y previsiones oceánicas producidas por sistemas de simulación y asimilación de los datos. Estos productos tienen una amplia gama de alcance espacio-temporal y resolución para atender a una variedad de posibles aplicaciones.

La oceanografía biológica representa el mayor percentual de investigadores en Brasil, notadamente porque las universidades y el cuerpo docente estudia, en su mayoría, lo referente a la interacción del hombre con los recursos vivos del mar, principalmente en la fase inicial de los cursos de oceanografía, lo que requiere más profesores, generando más investigaciones. En esta fase inicial, trata todo lo relacionado con los organismos que habitan en los mares: características, relación con el medio ambiente, comportamiento, etc.

En la secuencia, tenemos la oceanografía física com el segundo mayor índice de investigadores. Es una énfasis de la oceanografia que se ocupa de los movimientos de las aguas oceánicas con todos los fenómenos que las acompañan (oleaje, mareas, corrientes, etc.), así como de la relación del océano con la atmósfera.

En relación a la Oceanografia Geológica, ocupa el tercero puesto en relación a la cantidad de investigadores brasileños que se ocupan de esta área. Su campo de investigación es más específico que los dos anteriores, lo que explica la menor cantidad de datos primarios. Para recoger datos en esta área, la Oceanografia Geológica presenta estudio geológico de la superficie terrestre cubierta por el agua del mar, de las islas oceánicas y de las zonas costeras y entre otras cosas se ocupa del origen de los bordes continentales y de las formaciones geológicas con ellas relacionadas.

Por fin, la Oceanografia química en menor índice, representa la relación entre los componentes químicos del agua del mar con la abundancia de organismos, el intercambio entre el océano y la atmósfera y los efectos de la eliminación de desechos al mar. En Brasil es una área con recursos en la iniciativa privada y en algunos cursos Stricto Senso, lo que determina la menor cantidad de investigadores que generan datos primarios.

5.7.3.2 Características de los datos de investigación

En este escenario 97,7% de los investigadores brasileños respondieron la encuesta afirmando que trabajan con datos em formato digital y solamente 2,2% contestaran que no trabajan con formatos digitales.

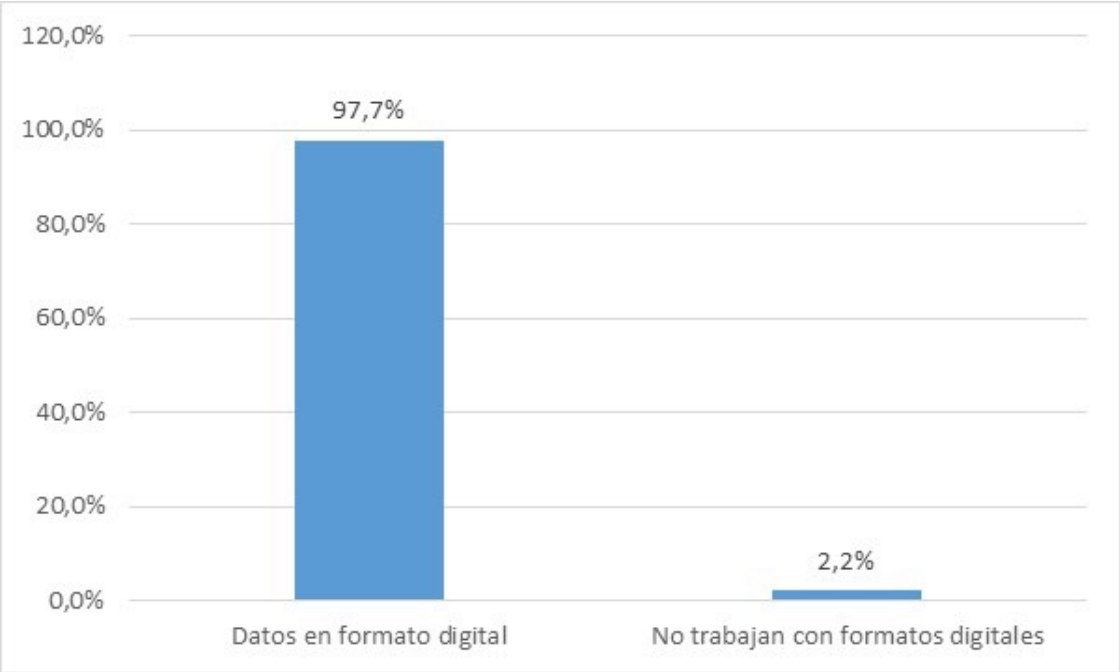

Gráfico 2: Características de los datos de investigación
Fuente: elaboración propia (2016)

Aunque las ciencias marinas se caractericen por la complejidad de su campo de estudio, apenas en los últimos años se han comenzado a fabricar los instrumentos y aparatos que permiten examinar la dinámica de las aguas del mar y observar sus cambios físicos y químicos, así como los efectos que éstos tienen en la superfície y las poblaciones de seres vivos. Mismo con la reciente inserción tecnológica en todos los ámbitos de los oceanográficos, el uso de formatos digitales es una realidad y un senso común entre los investigadores que responderan la encuesta. Todavía, algunos registros siguen en formatos que no son digitales, principalmente los datos observacionales, en algunos casos preservados solamente en ficheros y archivados en sectores (laboratorios, etc) o carpetas personales de los investigadores.

5.7.3.3 Tipología de los datos

En cuanto a la tipología de los datos, los resultados de la encuesta demuestran que el tipo de datos más frecuente que producen los investigadores son los Observacionales (61,4%), seguido por los Vídeos (50%) y Estadísticos (41,3%).

El índice del uso de imágenes coincide con el mesmo indice de respuestas de las Fotografias (40,9%), los Registros de salidas de campo (31,8%), el Audio (2,3%), encontrándose las Simulaciones (15,9%), siendo la opción Otros (13,6%) con el menor índice de respuestas.

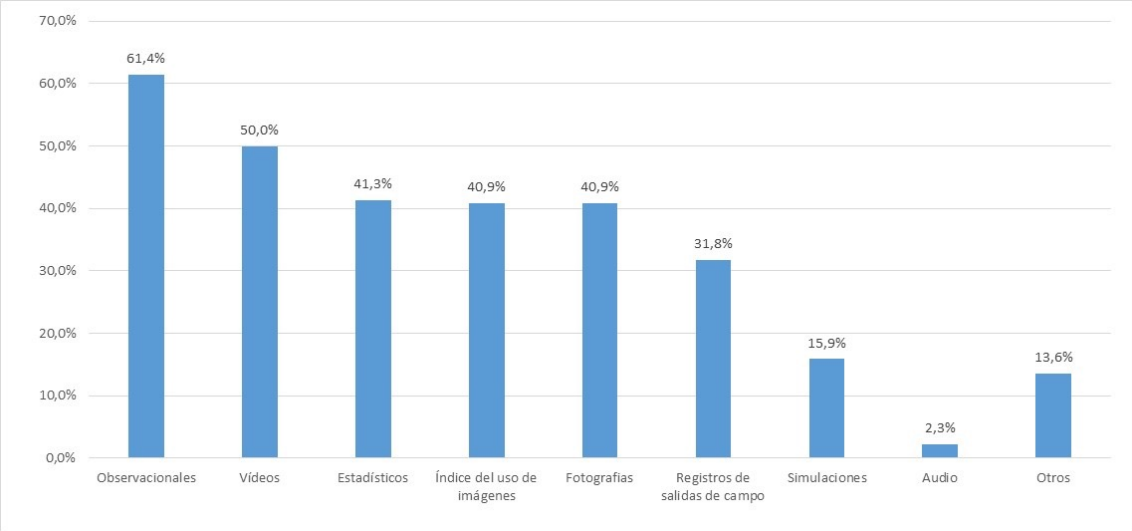

Gráfico 3: Tipología de los datos
Fuente: elaboración propia (2016)

La tipología de datos con presenta tamaña variedad, que es imposible que una sola ciencia los englobe. Por ello, el estudio actual de los océanos es realizado por un complejo de ciencias y tecnologías que, en su conjunto, constituye la oceanografía…

[…] Los datos de los sensores remotos son obtenidos en la Costa Atlántica y Antártida y distribuidos entre científicos de áreas de investigación comunes. Estos incluyen datos para monitorear el nivel de mar, la temperatura de la superficie del mar, el viento en la superficie, el hielo del mar, el color del océano y la salinidad de la superficie. Estos datos se hacen disponibles en varios niveles de procesamiento, cubriendo desde registros de datos sin procesar hasta productos geofísicos sintetizados tales como indicadores oceánicos y climatológicos.[…] Los datos in situ son obtenidos a través de varios sistemas y programas, incluyendo la captura de imágenes, muestras de sangre de animales marinos y el sistema de boyas para análisis en el océano […] proporcionan mediciones del estado interior del océano en la profundidad, lo que no es observable vía satélites de sensores remotos. **(Entrevistado 3)**

5.7.3.4 Formatos de los datos

En relación a los formatos digitales utilizados por los investigadores que respondieron la encuesta, para preservación de los datos em ambiente local (laptops, intranet, etc) los más utilizados son las Hojas de cálculo (70,5%), seguido del Texto (.txt o .doc) (50%). En la secuencia, los Formatos de imagen (45,5%) y los Sistema de Informação Geográfica (SIG) (38,6%), encontrándose Portable Document Format (.pdf) (29,5%). La alternativa Otros recibió 20,5% de las respuestas; Software estadística (15,9%) y Formatos de imagen vectorial (15,9%). Los Archivos de bases de datos (4,5%), encontrándose los formatos que solamente un investigador apuntó como alternativa: Rich Text Files (.rtf) (2,2%), Hypertext markup language (HTML) (2,2%) y Extensible markup language (XML) (2,2%). La opción Formatos de audio no fue apuntada por ningún investigador.

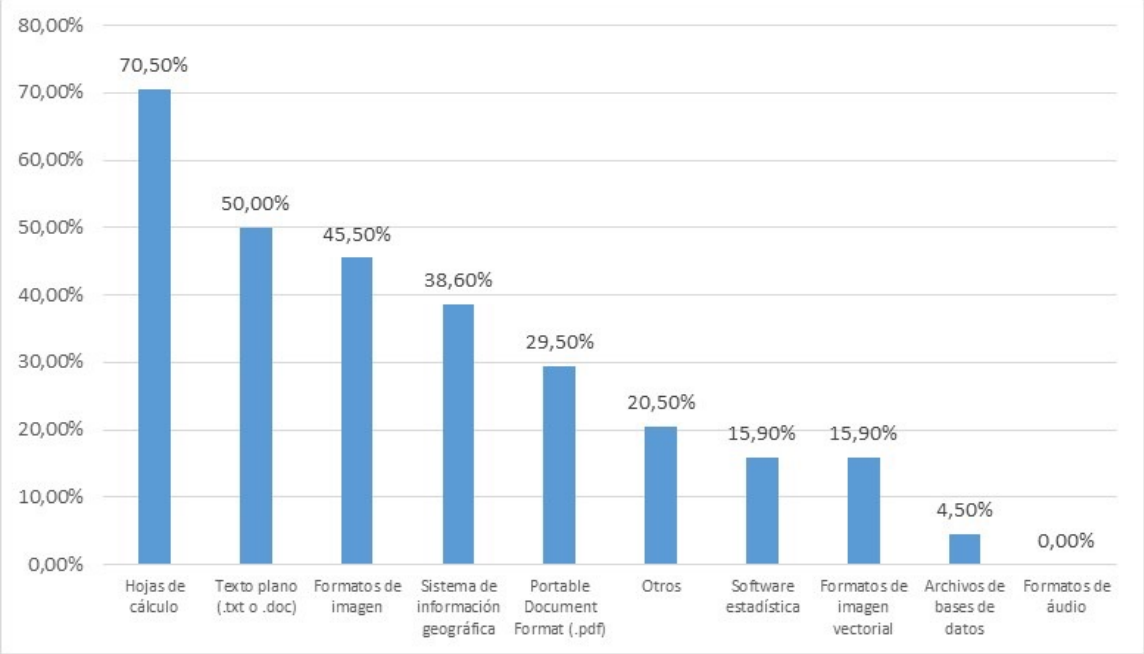

Gráfico 4: Formatos de los datos
Fuente: elaboración propia (2016)

Es importante para todos los investigadores entender el proceso que se ha llevado a cabo para generar un formato de datos determinado…

[…] El conocimiento de todo el contexto de un producto de datos implica la capacidad de rastreo de los eventos para una producción operativa y la comprensión de los atributos del producto (incluyendo lo que es producido, la

región oceánica cubierta y su resolución de escala, como ha sido producido, quien lo produjo, cuando y donde se ha hecho disponible, por cuanto tiempo, con que exactitud, el formato de entrega y servicios de entrega en red, política de servicios de datos y de la red, y más). Esta información permite al usuario decidir sobre idoneidad de un producto de datos para un propósito determinado. **(Entrevistado 1)**

5.7.3.5 Formatos de datos oceanográficos

Los formatos de datos oceanográficos incorporan componentes que impactan de una manera u otra en la preservación de datos. Por ejemplo, tecnologías o infraestructura de redes, interoperabilidad entre repositorios de datos, inovaciones en hardware, procesos o tecnologías generales como la virtualización y otras.

Antes que sofisticados sistemas de datos puedan ser construidos, debe haber una adopción extendida de enfoques/estrategias comunes para describir y transportar datos oceanográficos. De esta manera, los formatos de datos oceanográficos tratan del uso de formatos habitualmente utilizados por los investigadores para archivar los datos que generan.

De acuerdo con las respuestas obtenidas en la encuesta, existen siete formatos digitales utilizados para crear o registrar datos primarios que juegan un gran papel en la armonización de los datos oceanográficos en Brasil: XML (32,6%), seguido por el Shapefiles (30,4%) y el CSV (26,1%) y el GeoTIFF (26,1%). La opción Otros (26,1%), encontrándose NetCdf (19,6%) y Matlab (21,7%), seguido por ASCII (15,2%); KML (6,5%) y el HDF (4,3%). Los formatos GRIB, Audio y JSON (0%) no tuvieron respuestas.

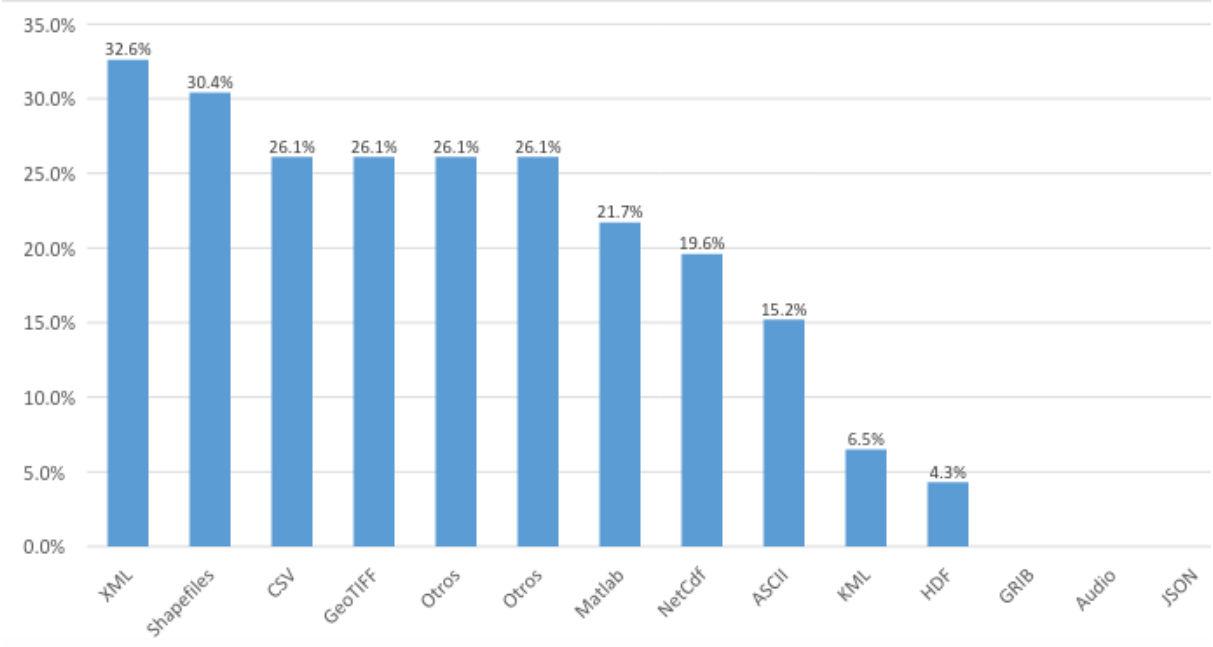

Gráfico 5: Formatos de datos oceanográficos
Fuente: elaboración propia (2016)

[…] Generalmente los datos son archivados en repositorios o bases de datos locales o distribuidos en repositorios de datos internacionales. El Banco Nacional de Datos Oceanográficos, (www.bndo.gov.br) también recibe datos de diferentes redes, en coherentes grupos de datos para propósitos de oceanografía operacional (asimilación, validación). Los datos pueden ser perfiles de medición individuales o mapas sintetizados de observaciones globales. **(Entrevistado 1)**

[…] El conocimiento de todo el contexto de un producto de datos implica la capacidad de rastreo de los eventos para una producción operativa y la comprensión de los atributos del producto (incluyendo lo que es producido, la región oceánica cubierta y su resolución de escala, como ha sido producido, quien lo produjo, cuando y donde se ha hecho disponible, por cuanto tiempo, con que exactitud, el formato de entrega y servicios de entrega en red, política de servicios de datos y de la red, y más). Esta información permite al usuario decidir sobre idoneidad de un producto de datos para un propósito determinado.

[…] Aunque un gran número de formatos de archivo estén disponibles para expresar datos oceanográficos, desde forma libre, archivos de texto (el ASCII), a formatos binarios altamente estructurados. Estos formatos a menudo difieren a un nivel fundamental, haciendo difícil desarrollar herramientas y aplicaciones que trabajen con todos los formatos. […] Para paliar esta dificultad, la

comunidad oceanográfica se ha estandarizado en torno a NetCdf[34], que proporciona un formato de archivo orientado a la matriz, un formato de archivo independiente de la plataforma que puede contener una amplia variedad de tipos de datos, desde mediciones in situ, a grandes tablas de datos multidimensionales de modelos numéricos. **(Entrevistado 3)**

[…] El formato de NetCdf es respaldado por bibliotecas de software de alta calidad, en una variedad de idiomas, que facilitan el proceso de desarrollar aplicaciones que consumen y producen datos en este formato. Además, algunas de estas bibliotecas de software (p.ej., la biblioteca oficial Java de NetCdf) son capaces de leer una variedad de otros formatos de archivo (como el formato Binario GRIB-GRIdded, un estándar de la Organización Mundial Meteorológica para codificar datos de previsión) y los interpreta como si fueran archivos NetCdf. De este modo, la comunidad oceanográfica brasileña ha conseguido armonizar conjuntos de datos anteriormente dispares. **(Entrevistado 2)**

5.7.3.6 Metadatos

En relación a la normalización de metadatos, los investigadores respondieron que utilizan: Muestra CSR Record y la opción Otros (38,6%), seguido de la ISO 19115 - Informação geográfica (22,7%) y el Resumen de relatorios de cruceros (22,7%). En menor indice de uso, el Common Data Index (11,4%) y el Geonetwork (11,4%), encontrándose el Marine Community Profile (2,3%) y el Cruise Summary Report (CSR) (2,3%). Por fin, los que no fueron indicados por ningún investigador, el Directorio Interchange Form (0%) y la Base de datos CSR/ROSCOP (0%).

---

34 Nota del autor: la comunidad oceanográfica acompaña notícias sobre el formato NetCdf en el sitio http://www.unidata.u-car.edu/software/NetCdf

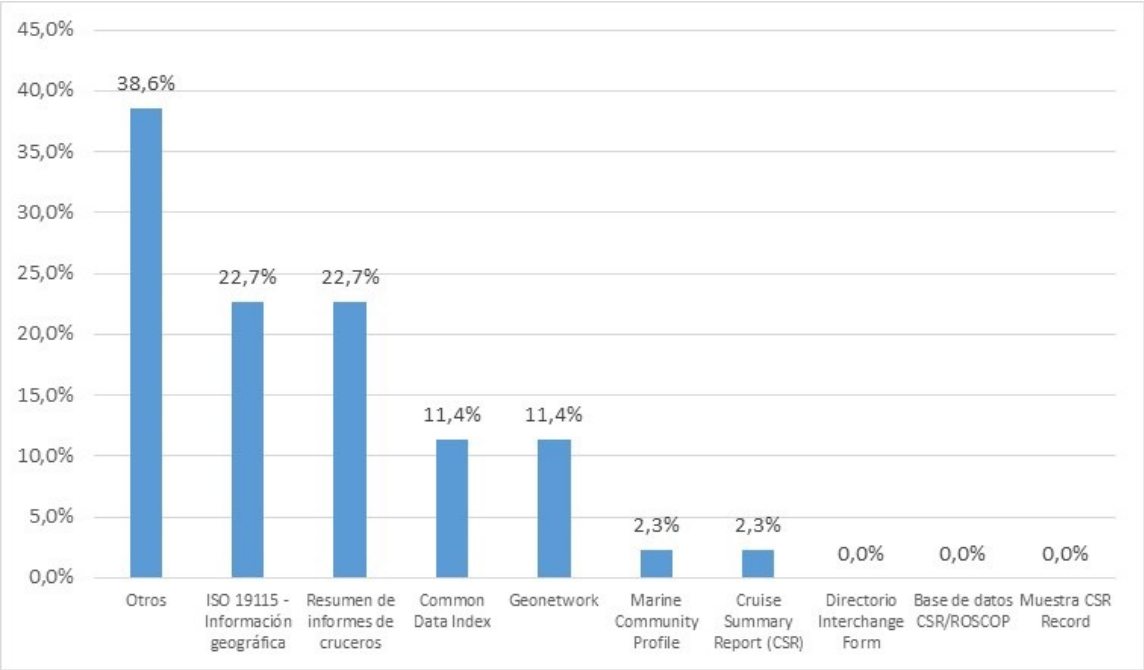

Gráfico 6: Metadatos
Fuente: elaboración propia (2016)

La no existencia de un estándar único para la preservación de datos oceanográficos que siga una ontología determinada se convierte en una dificultad añadida. El que algunas instituciones utilicen sus propios estándares (ej. SeaDataNet, SISMER, etc) dificulta la integración de los metadatos. A menudo los editores de metadatos son demasiado complejos para ser utilizados en entornos marinos. Es por ello que se han desarrollado unas aplicaciones web que hacen más sencilla la creación de los metadatos, bajo los formatos de metadatos que uno u otro centro de investigación brasileño utiliza. Dichos metadatos son integrados usando tecnología que permite cumplir con los estándares de la ISO19115. Por otro lado, es curioso que el GeoNetwork aún no sea muy utilizado en el escenario brasileño y que sea notable que la opción Otros demuestra la dispersión existente en los formatos de metadatos. Siendo una aplicación de software libre y código abierto, diseñada para permitir acceso a bases de datos georreferenciadas y metadatos de varias fuentes, el Geonetwork tiene capacidad para mejorar el intercambio de datos entre los centros de investigación de Brasil, además de facilitar el intercambio con los principales repositorios internacionales. Debido a que el Geonetwork utiliza el protocolo Z39.50, puede acceder a catálogos remotos y hace que sus datos estén disponibles para otros servicios de catálogo.

5.7.3.7 Aplicación de software

En relación a las aplicaciones de software utilizadas para gerenciar los dados, los investigadores que respondieron la encuesta apuntaran: Software para tratamiento de datos cuantitativos (84,1%), seguido de Software para edición de imágenes (37%) y Software para tratamiento de datos cualitativos (34,8%). La opción Otros (15,2%) y la opción Software para transcripción de audio y vídeo (4,3%).

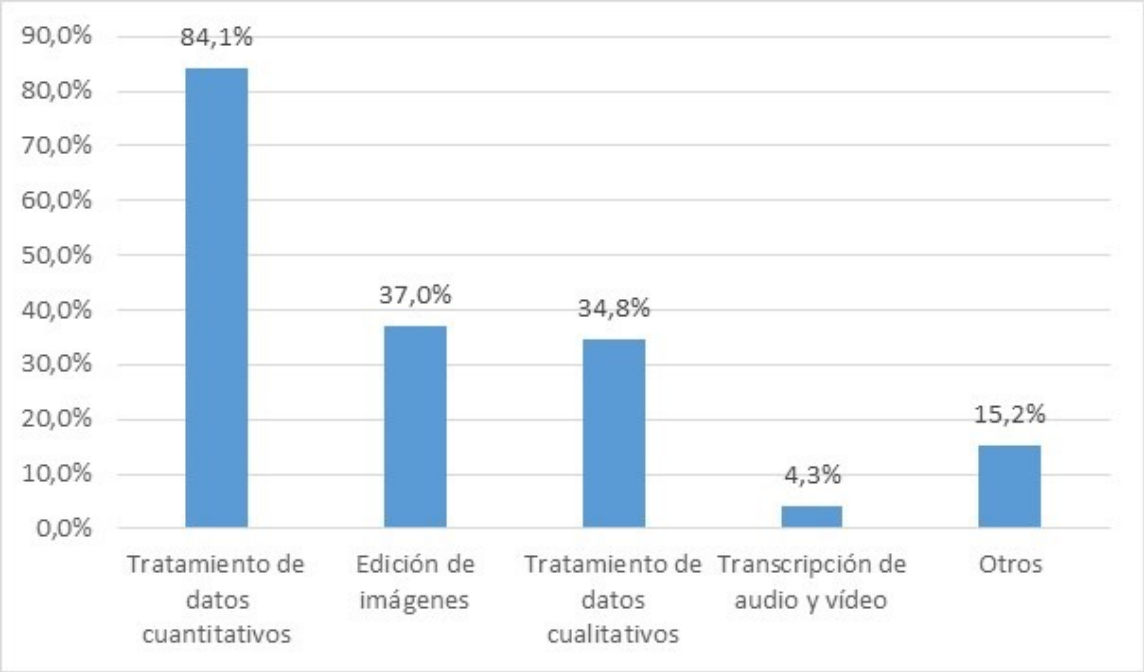

Gráfico 7: Aplicación de software
Fuente: elaboración propia (2016)

Un aspecto esencial de la coordinación de datos y acciones de interoperabilidad por medio del uso de softwares adecuados se refiere a la necesidad de crear normas y estándares semánticos que establecen una base común para diferentes proyectos, y juntos los aspectos de diferentes disciplinas y ramas del conocimiento bajo un mismo paraguas conceptual.

[…] Una Infraestructura de Datos Espaciales (IDE) es compuesta de políticas, tecnologías y estándares para interconectar a los usuarios de información espacial. Por lo tanto, la implementación de un programa de IDE debe aspirar a la organización de la información geográfica, lo que resulta en el desarrollo de

normas de gestión centralizada y distribución jerárquica, a nivel local, regional, nacional y supranacional. **(Entrevistado 3)**

Ejemplos internacionales de la implementación exitosa de la IDE son notorios, especialmente para datos marinos[35]. En Brasil, los esfuerzos a nivel nacional se han organizado desde el proceso de implementación del programa de INDE/ Brasil (Infraestructura Nacional de Datos Espaciales; http://www.inde.gov.br/), coordinado por la Comisión Nacional Cartografía (CONCAR), cuyo objetivo es la gestión de pedidos y el almacenamiento y el acceso de los datos geoespaciales y metadatos a nivel nacional.

[…] *Sin embargo, bajo el INDE, no existen directrices específicas relativas a los datos marinos o costeros. Es importante destacar que el desarrollo de una IDE no debe limitarse a los aspectos técnicos y operativos relacionados con las tecnologías de gestión de datos, sino también las acciones políticas y administrativas que proporcionan una amplia cobertura y difusión de su esencia y de la documentación (metadatos)*. **(Entrevistado 2)**

[…] El formato NetCdf proporciona una forma de datos simple y de disciplina neutra, para codificar matrices multidimensionales y sus atributos. Las convenciones CF[36] proporcionan la semántica adicional definiendo como codificar datos oceanográficos (y datos de otras disciplinas) en archivos NetCdf. […] Estas convenciones están actualmente enfocadas a la descripción de tablas de datos de modelos numéricos o productos satelitales analizados. Ellas proporcionan el medio para describir la tabla sobre la cual los datos están expresados, junto con una suite de "nombres estándar", que son usados para identificar la cantidad geofísica que estos datos representan[37]. **(Entrevistado 3)**

[…] *El formato de archivo NetCdf y las convenciones CF (también conocido en Brasil como CF-NetCdf) proporcionan un medio eficaz para codificar datos oceanográficos. Numerosas herramientas para el análisis y la visualización de datos oceanográficos han sido desarrollados sobre la base de estas*

---

[35] **Nota del autor**: Se destacan en este contexto la INSPIRE (Infraestructura de información espacial en la Comunidad Europea; http://inspire.jrc.ec.europa.eu), que sirvió de base para el programa específico de EMODNET datos marinos Europea (European Marina Observación y red de datos, www.emodnet-biology.eu)

[36] http://www.cfconventions.org

[37] Nota del autor: p.ej.,"temperatura_potencial_agua_marina"

tecnologías, incluyendo herramientas de escritorio [38]y herramientas basadas en Webs[39]*. […] Los archivos resultantes son totalmente compatibles con CF y pueden ser leídos por numerosas aplicaciones genéricas compatibles con CF. Nótese, sin embargo, que tales archivos pueden describir sólo un resultado de observación tales como una sola serie temporal, una reseña, o un rastreo de barco. […] No extensamente consensuado de acuerdo a las normas existentes, pues aun ha de describir colecciones de observaciones y esto es un obstáculo clave para el desarrollo de sistemas que permitan a los usuarios visualizar y procesar datos in situ. La última versión de los NetCdf*[40] *contiene nuevas características que lo hacen apropiado para codificar tales colecciones de observaciones y la investigación está en desarrollo hacia como puede alcanzarse esto en la práctica*. **(Entrevistado 2)**

Mientras NetCdf proporciona un formato de datos constante para almacenar datos oceanográficos y los investigadores utilizan una consistente descripción de meta información de los datos que almacenan, los repositorios internacionales ofrecen mecanismos con el cual los datos pueden ser accedidos desde Internet.

[…] Específicamente, los investigadores pueden acceder a los subconjuntos de conjuntos de datos de otros científicos y registran directamente sus paquetes de análisis. Ellos pueden también ofrecer integración de grandes grupos de tablas de datos que se encuentren en varios archivos. Estas capacidades son importantes, porque un científico que desee acceder a un conjunto de datos oceanográficos (por ejemplo, un reanálisis multidecadal del océano) a menudo no necesita el conjunto completo de datos, que pueden tener cientos de gigabytes o incluso terabytes de tamaño. Además, a menudo no es deseable considerar el conjunto de datos como un gran grupo de archivos individuales: el científico puede preferir considerarlo como un gran conjunto de datos

---

[38] p.ej., Ferret [http://ferret.wrc.noaa.gov/], CDAT (http://cdat.sf.net), GrADS (http://www.iges.org/grads/), Ocean Data View(http://odv.awi.de/), e Ingrid (http://iridl.ldeo.columbia.edu/dochelp/Documentation/)

[39] (p.ej., Live Access Server [ver la discusión posterior; Schweitzer et al., 2007], Godiva2 [Blower et al., 2009], y DChart (http://www.epic.noaa.gov/epic/software/dchart/).

[40] **Nota del autor**: la versión 4

cuadridimensional, que puede ser sub-muestreado de numerosos maneras. **(Entrevistado 2)**

[…] Hay una relación muy cercana entre el formato de archivo NetCdf y los formatos particulares que los investigadores almacenan sus registros. La diferencia que es posible transmitir datos NetCdf con casi ninguna pérdida de información. **(Entrevistado 3)**

Muchas herramientas de análisis de datos de escritorio tratan a los datos NetCdf almacenados localmente exactamente de la misma manera que la los datos almacenados remotamente en un servidor, proporcionando a los científicos la capacidad de analizar y visualizar enormes cantidades de datos repartidos.

[…] Los servidores pueden actuar como medio para acceder a los datos que están alojados por centros de datos en muchos otros formatos de archivo como HDF, GRIB y BUFR. Estos formatos son populares en otras comunidades tales como observatorios de la Tierra y la meteorología. […] El usuario final no necesita saber nada acerca de que formato de datos se utiliza en un servidor remoto ya que la mayoría de los repositorios atienden normas internacionales, lo que torna posible una muy ponderosa tecnología para la armonización e integración de datos. **(Entrevistado 2)**

Cada proveedor de datos oceanográficos ha implementado un portal Web para usuarios, para descubrir y explorar sus productos, y para proporcionar a estos usuarios enlaces para descargarlos. Existen catálogos dedicados en muchos de estos sitios para ayudar a los usuarios a descubrir conjuntos de datos específicos. Esta estructura ha conducido al desarrollo de un gran número de portales Web - cada uno está diseñado de manera diferente, lo que a menudo resulta confuso para los usuarios, particularmente para aquellos ajenos a la comunidad oceanográfica.

[…] hay esfuerzos en curso para crear catálogos integrados que provean a los usuarios de un solo punto de encuentro, para una integración de productos de datos contenidos en repositorios de datos, pero este desarrollo aún está centralizado en áreas aisladas y no si puede decir que representan el escenario mas amplio de la oceanografía brasileña. **(Entrevistado 2)**

5.7.3.8 Alternativas para compartir los datos

En relación la manera como otros investigadores pueden obtener acceso a los datos de sus investigaciones, la encuesta demuestra que: Datos distribuidos por email (58,7%); Dispositivos de armazenamento portáteis (Ex: CD, Pen Drive, etc.) (54,3%); en formato impreso (28,3%); Realizan download para un servidor web y disponibilizan acceso por medio de solicitud (26,1%); Realizan download para un servidor web de acesso público (Ex: repositórios institucionales) (15,2%); la opción Otros (15,2%); Dejan todos los datos relacionados a um articulo publicado sob responsabilidad y decisión del editor de una revista (6,5%); No es posible acceder a los datos de sus investigaciones (2,2%).

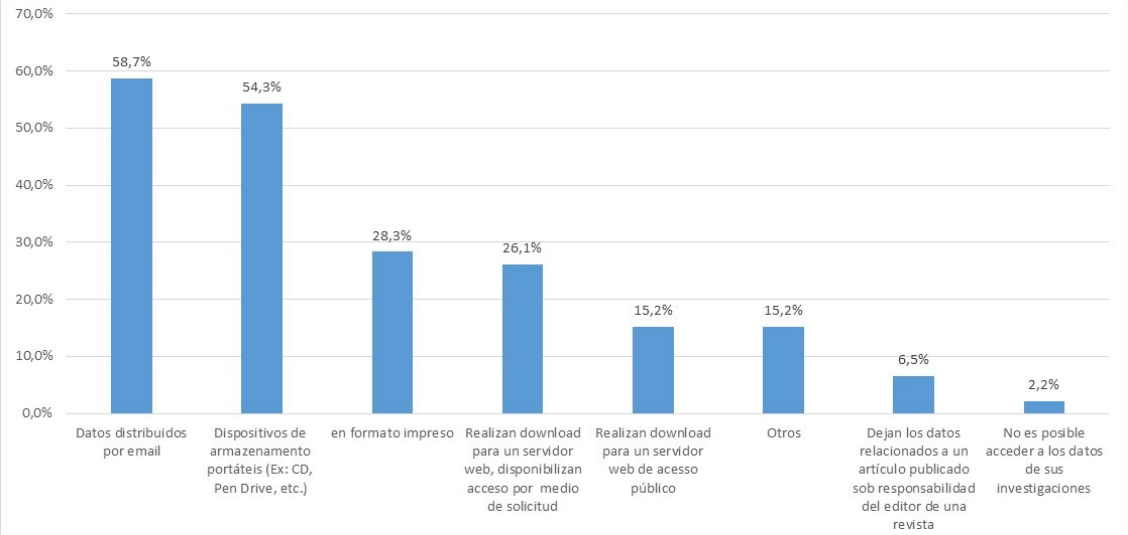

Gráfico 8: Alternativas para compartir los datos
Fuente: elaboración propia (2016)

Habiendo encontrado datos de potencial interés a través de una búsqueda de texto en uno de los sitios Web, el usuario a menudo querrá evaluar la aptitud de estos datos para su aplicación antes de adquirirlos. Muchos servicios de visualización están ahora implementados, proporcionando acceso a visualizaciones predefinidas o generadas dinámicamente.

[…] Los estándares de datos de bajo nivel descriptos en la sección sobre tecnologías de base, son aquí sumamente importantes: no sería factible proporcionar acceso visual a todos los diferentes grupos de datos en los repositorios de datos sin acordar previamente como los datos son formateados. **(Entrevistado 2)**

Los datos oceanográficos compartidos en el escenario brasileño, es decir, los servicios de visualización primarios, fueron definidos para permitir la visualización de gráficos predefinidos actualizados diariamente.

[…] *Estos servicios son fáciles de usar, eficientes y rápidos. Las imágenes están previamente preparadas, basadas sobre criterios cuidadosamente preseleccionados, pero tales servicios proporcionan al usuario de una limitada o nula capacidad para ajustar los gráficos (p.ej., cambiar la escala de color o la región de interés).* **(Entrevistado 1)**

5.7.3.9 Repositorios

En relación a los repositorios en que los investigadores brasileños que respondieron la encuesta depositan los datos de sus investigaciones, la siguiente tabla demuestra la cantidad de investigadores que han contestado cada alternativa. El resultado apunta el siguiente:

REPOSITORIO	Consultar	Upload de los datos	Download de los datos
Australian Ocean Data Center Facility (AODC)	0	1	3
Australian Ocean Data Center Facility (AODC)	0	1	3
British Oceanographic Data Centre (BODC)	1	1	1
European Marine Observation and Data Network (EMODnet)	0	1	1
Geological and Geophysical Data (Geo-Seas)	3	0	0
Integrated Marine Observing System (IMOS)	1	0	5
Intergovernmental Oceanographic Commission	1	1	6
JERICO	0	0	1
National Oceanic and Atmospheric Administration (NOAA)	3	0	7
Rolling Deck to Repository	0	1	0
SeaDataNet	1	1	0

REPOSITORIO	Consultar	Upload de los datos	Download de los datos
Systèmes d'Informations Scientifiques pour la MER (SISMER)	1	0	1

Tabla 18: Repositorios utilizados
Fuente: elaboración propia (2016)

El resultado claramente demuestra la falta de preferencia por parte de la mayoría de los investigadores brasileños en un mismo repositorio. Aunque existan repositorios de datos más consolidados entre la comunidad oceanográfica internacional, esta claro que para los investigadores brasileños esto no es un factor determinante para escoger el repositorio. En general, depositan los datos de investigación donde la área que trabajan sea más compatible con las temáticas de sus proyectos, o mismo cuando perciban facilidad operacional para realizar el deposito.

Para atender las crecientes demandas y diminuir las dificuldades de los investigadores en el momento de hacer el deposito de los datos, una nueva generación de sistemas de visualización basados en Web ha surgido recientemente, basada sobre conceptos y estándares de la comunidad de Open Geographic Information Systems (OpenGIS). Son recursos que permiten superponer diferentes fuentes de información: por ejemplo, un usuario puede superponer una previsión de altura de la superficie del mar sobre un mapa de densidad demográfica para obtener un rápido panorama de los riesgos que puede plantear la predicción de una marea tormentosa. Esta tecnología ha sido demostrada junto al proyecto oceanográfico Europeo MERSEA (http://www.resc.reading.ac.uk/mersea) y ECOOP (http://www.resc.reading.ac.uk/ecoop) y han sido adaptados para crear sistemas de Vista Rápida Dinámica (Dynamic Quick View (DQV).

Los bancos/fuentes de datos oceanográficos son muy extensos: por ejemplo, los sistemas de predicción de corrientes oceánicas, estudios ambientales y el ecosistema marino generan enormes cantidades de datos oceanográficos por año, por cada organización.

[…] Se espera que esta cantidad se incremente rápidamente con el despliegue de nuevos sistemas de observación y con los aumentos en la resolución y la

*complejidad de los modelos numéricos. No hay ninguna autoridad centralizada con los recursos para supervisar la dirección de esta extensa colección. Además, transferir toda esta información a una ubicación central sería poco práctico y forzaría a los proveedores de datos a abandonar parte del control sobre su propia producción de datos. Así, un acercamiento de gestión de datos repartido, que proporcione la capacidad de compartir datos oceanográficos eficazmente a través de Internet, es esencial al propósito brasileño de desarrollar un sistema de pronóstico oceanográfico global. […] Además de solucionar muchos de los problemas asociados con el manejo de grandes bancos de datos, un sistema repartido puede proveer a los usuarios con servicios de datos más confiables y eficientes por la reproducción de datos, así como de un uso más eficiente de redes. (***Entrevistado 1)**

El acercamiento esencial a repositorios de datos oceanográficos es una suite de herramientas basadas en estrategias compartidas. Estas herramientas han sido diseñadas principalmente para ser utilizadas por la comunidad científica, pero proporcionan una sólida base sobre la cual otros sistemas pueden ser construidos para servir a otras comunidades de usuarios tales como consorcios y acuerdos de cooperación internacionales. En el caso del Tratado Antártico, Brasil aún no dispone un repositorio de datos accesible, dejando abierta su participación en proyectos consolidados como el Antarctic Master Directory.

[…] Estas herramientas permiten a los científicos usar datos en un modo que los libera de la necesidad para entender los detalles de bajo nivel de formatos de archivo, estructura, o incluso de la ubicación física de los datos. La idea de ocultar esta complejidad es fundamental para el éxito de un sistema de datos que debe tratar con tal gran variedad y volumen de información. […] Esta complejidad oculta debe estar equilibrada con la necesidad de soportar una muy amplia gama de aplicaciones, y esto requiere flexibilidad.

[…] *Esta integración con la tecnología GIS representa el comienzo de una importante nueva dirección en el análisis de datos oceanográfico y su visualización. Esto tiene el potencial de permitir combinar datos oceanográficos con muchas otras fuentes de datos provenientes de otras comunidades, dando soporte a una amplia gama de nuevas aplicaciones en la ciencia y en la toma de decisiones.* […] *Un reciente avance en la visualización de datos oceanográficos*

es el desarrollo de sistemas de visualización interactivos basados en Web. Estos sistemas permiten al usuario interactuar directamente con los datos y personalizar la visualización en cierta medida. **(Entrevistado 3)**

Los resultados de la encuesta y las entrevistas demuestran que los servicios de disponibilización de los datos encuentran este equilibrio proporcionando dos caminos diferentes para que los científicos accedan y usen los datos: (1) a través de portales Web, que ocultan la complejidad de los datos, pero proporcionan una fija gama de funcionalidad, y (2) permitiendo a los datos ser procesados por la herramienta de escritorio que el científico escoja. Estos dos enfoques pueden operar sincronizadamente, con el científico realizando el descubrimiento preliminar, la evaluación, y el análisis sobre el Web, para *a posteriori* y si fuese requerido, utilizar herramientas especializadas para llevar a cabo un trabajo más a fondo.

Este análisis resume los avances tecnológicos que se han llevado a cabo en el contexto de los datos oceanográficos en el escenario brasileño, facilitando la capacidad del investigador para recoger, evaluar, visualizar, descargar y analizar los datos oceanográficos.

Se han desarrollado muchos proyectos de investigación y datos de mapas marinos de la plataforma continental de Brasil en las últimas décadas, lo que genera una gran cantidad de información, con una significativa parte publicada en artículos científicos, pero sólo una pequeña cantidad de ella fue incluida en las bases de datos que sirven como plataformas para la difusión del conocimiento.

[…] *Estas plataformas son la base para meta-análisis que buscan estándares inexplorados, entre ellos los relacionados con la sinergia entre los datos de diferentes naturalezas. Este escenario se explica en parte por la ausencia de esfuerzos coordinados entre universidades/centros de investigación, organismos gubernamentales y organizaciones no gubernamentales que tienen como objetivo, o praxis, estructuración y normalización de los datos generados en estos estudios. Por lo tanto, la realidad actual muestra un aluvión de información desconectada, una pérdida sustancial de datos primarios y la imposibilidad de manejo y acceso a la información. La pérdida científica,*

logística y financiera de esta realidad es inmensa. En este sentido el escenario brasileño es antagónico a las tendencias internacionales que han alentado la organización de programas y la difusión de información como base para la investigación científica, en particular en lo que respecta a la difusión de datos de carácter ambiental. […] El desarrollo de estos programas no se limita a inventarios, pero se refiere a la creación de una nueva mentalidad en relación con la necesidad de compartir datos, sean primarios o procesados. **(Entrevistado 1)**

5.7.3.10 Factores motivacionales para compartir datos

La encuesta demuestra cuales factores son considerados motivacionales para compartir los datos científicos de las investigaciones en un repositorio digital: Beneficios para comunidad científica (80,4%), seguido de la alternativa Permitir colaboraciones y contribuciones de otros investigadores (56,5%), en seguida Beneficios potenciales para sociedad en general (50%); Aumentar la visibilidad de mis investigaciones (43,5%); Beneficios para institución en que trabajo (28,3%); Exponer los resultados obtenidos para validación por otros investigadores de área (26,1%); Requisito por parte de la agencia que investiga mi investigación (21,7%); Benefícios para mi carrera profesional (méritos para mi evaluación institucional, etc) (17,4%) y la opción Otros (2,2%).

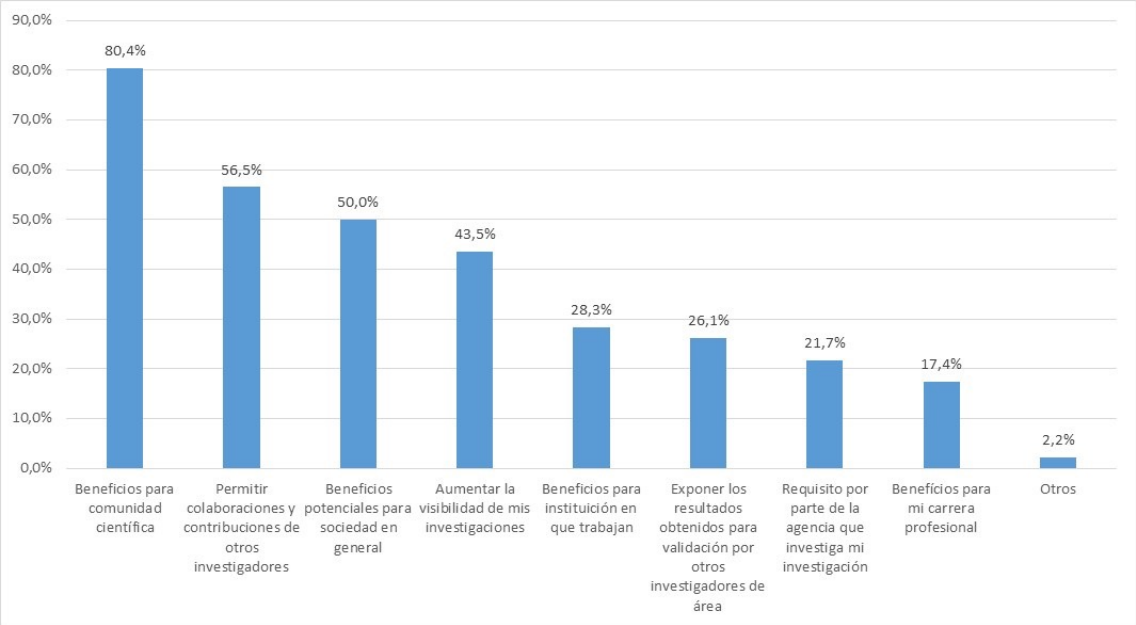

Gráfico 9: Factores motivacionales para compartir datos
Fuente: elaboración propia (2016)

Las respuestas a respeto de los factores motivacionales demuestran que en general los investigadores tienen interés en compartir los datos para el avance de la ciencia o mismo para promover la institución en que trabajan, además de obtener reconocimiento académico. Por otro lado, los factores desmotivacionales, conforme puede ser verificado en relación a los factores desmotivacionales para compartir los datos, apuntan preocupación con la seguridad sobre preservación de lo datos y desconocimiento de los investigadores sobre los procedimientos necesarios para compartir los datos.

5.7.3.11 Factores desmotivacionales para compartir los datos

En relación a los factores desmotivacionales para compartir los datos levantados en sus investigaciones, las respuestas apuntan el siguiente: Tiempo y esfuerzo necesario para compartir (47,8%); La falta de financiamiento en la área que desarrollo investigaciones (43,5%); Restricciones relativas a la seguridad y confidencialidad de los datos (43,5%); Desconocimiento de las normas y procedimientos para compartir mis datos (45,7%); Restricciones éticas (19,6%) y la opción Otros (17,4%).

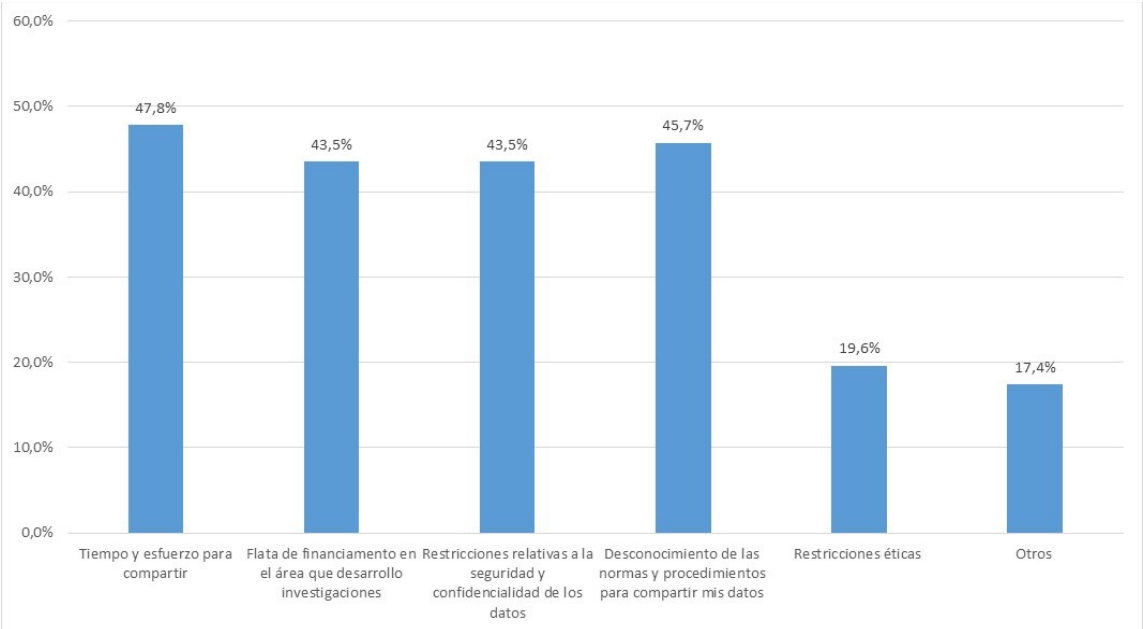

Gráfico 10: Factores desmotivacionales para compartir datos
Fuente: elaboración propia (2016)

A menudo, los investigadores mantienen los datos después de que el proyecto esté terminado en el mismo lugar en que se almacenaron durante el proceso de investigación, sin tener en cuenta la forma en que podrían afectar a su capacidad de uso y a su accesibilidad a largo plazo. Archivar datos conlleva la preservación activa de los datos, así como la adopción de medidas para aumentar su capacidad de descubrimiento y de accesibilidad. Se trata, entre otras cosas, de dar códigos identificadores únicos a los datos y a la realización de los controles comunes para su replicación.

La documentación adecuada es uno de los requisitos más importantes para reutilizar conjuntos de datos. Por ejemplo, es difícil para el investigador utilizar un conjunto de datos si no es capaz de determinar el significado de los nombres de las variables. Si los investigadores almacenan los datos digitales en servidores o discos duros sin realizar periódicamente las acciones de preservación necesarias, con el tiempo sus datos se tornarán inutilizables. También se debe saber quién posee y controla los datos y si hay problemas de privacidad. En muchos casos, los investigadores no tienen respuestas fáciles a estas preguntas, aunque estén cientes del prejuicio que pueden tener con malas prácticas de preservación o en peor hipótesis, ninguna acción para compartir los datos que tienen.

El desconocimiento de las normas y procedimientos para compartir los datos es una barrera que en el caso brasileño puede ser traspasada con algún esfuerzo. Por ejemplo, el Integrated Ocean Observing System (IOOS) enumera una amplia gama de recursos marinos en la web y cuenta con un portal de búsqueda para ayudar a encontrar y depositar los datos específicos de las investigaciones sobre océanos. Estos índices no necesariamente recogen datos, sino que apuntan a una serie de recursos sobre un tema en particular conjuntamente con bases de datos que también pueden estar disponibles en las bibliotecas.

5.7.3.12 Servicios de apoyo

Muchas herramientas de software dedicadas han sido desarrolladas y puestas a disposición para varias aplicaciones científicas. Uno de los objetivos primarios

de desarrollar tecnologías para la descripción, la visualización y el acceso a los datos, es posibilitar la interoperabilidad entre diferentes proveedores e/o repositorios de datos.

En relación a los servicios de apoyo, a respeto de la pregunta "¿Estás interesado en tener servicio de asistencia para compartir tus datos científicos?", los investigadores contestaron lo siguiente: Asesoramiento para almacenar, gestionar y compartir de manera segura (65,2%); Asesoramiento en cuestiones prácticas con la gestión de datos (43,5%); Asesoramiento para la preservación de los datos recogidos en mi investigación (41,3%); Asesoramiento en la creación de un plan de gestión de datos científicos (41,3%) y la Opción Otros (19,6%).

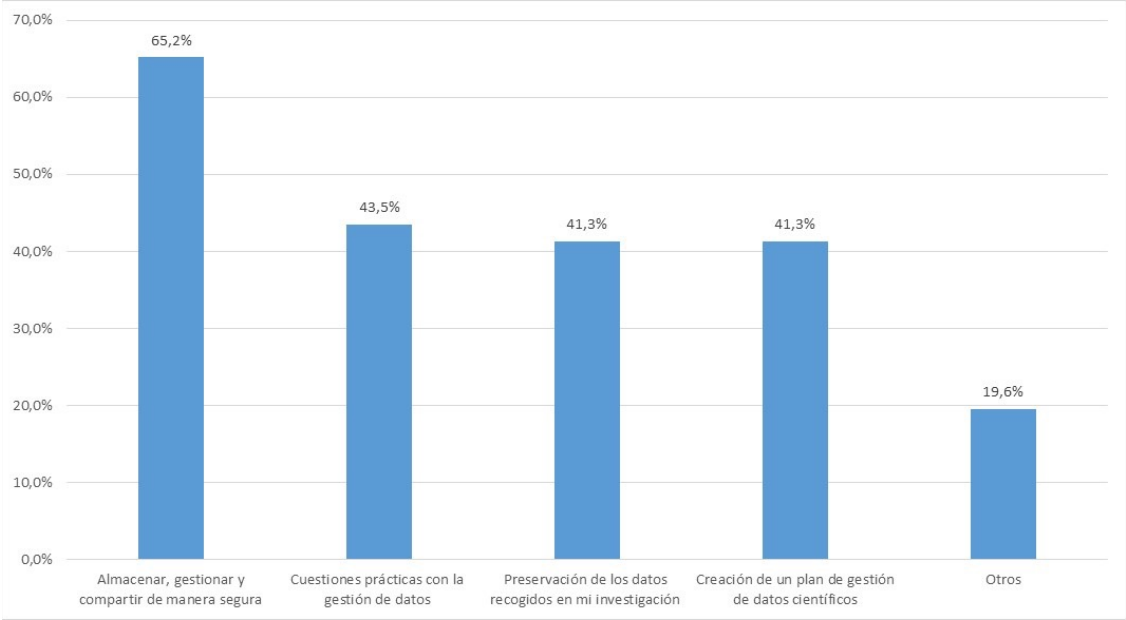

Gráfico 11: Servicios de apoyo
Fuente: elaboración propia (2016)

[…] El análisis oceánico y los centros de previsiones producen datos a escala global o regional, proporcionando previsiones en tiempo real (diario y estacional) y series histórico temporales. Los modelos numéricos resuelven una variedad de características oceánicas, desde remolinos, a la circulación a gran escala global; pueden ser análisis retrospectivos, análisis (la mejor estimación del estado actual), o previsiones. **(Entrevistado 1)**

En el Espacio para comentarios adicionales de la encuesta tres investigadores respondieran:

[…] La Marina de Brasil tiene una base de datos oceanográficos. Sin embargo, tiene los mismos problemas que tenemos. La falta de inversión para continuar el banco. La mayoría de los datos recogidos por las instituciones de investigación no se envían a la Dirección de Hidrografía y Navegación. Del mismo modo la Sociedad de Investigación de Recursos Minerales (CPRM) actual Servicio Geológico de Brasil encargada de gestionar los datos de la oceanografía geológica/geología marina no recibe información y no invertir en ella juntos universidades. Creo que la falta de una política de implementación de almacenamiento de datos en un banco de esta información que creo crucial para el país.

[…] El depósito de los datos es realizado en mi laboratorio en la base de datos de OBIS y NonatoBase, gestionados por los investigadores de la UFSC. Estos datos también alimentaron las bases de datos de Petrobras. La plataforma para compartir los datos está en la etapa final de montaje y está siendo operada por consulta individual en este momento.

[…] Utilizo los datos del Banco Nacional de Datos Oceanográficos (BNDO) y OceanColor

[…] *En Brasil, algunas herramientas fueron puestas en práctica de manera integrada via web por universidades de manera similar a las implementaciones realizadas en Europa, Estados Unidos y Australia, utilizando el servidores institucionales de los participantes; pero esta integración no siguió adelante y actualmente la mayoría tiene su propia manera de compartir los datos .* **(Entrevistado 3)**

[…] *La profusión de diferentes catálogos y portales Web para acceder a datos oceanográficos puede aumentar la dificultad del usuario para descubrir y acceder a los datos que necesita. Será muy importante para los proveedores de datos el presentar sus catálogos de información en una forma que puede ser recolectada y agregada por "los meta-catálogos" que pueden actuar como los puntos únicos de primer contacto para nuevos usuarios. Estándares*

internacionales están surgiendo, lo que ayudará a posibilitar el desarrollo de estos meta-catálogos. **(Entrevistado 2)**

5.8 Valoración de la situación

Los desafíos del desarrollo para la integración y la sistematización de los datos oceanográficos en Brasil indican algunas cuestiones clave que aún dificultan la adopción e implementación de grandes programas de gestión de datos.

El primer problema es la ausencia de una infraestructura adecuada en el ámbito nacional para recoger los datos de investigación oceanográfica, generando muchas veces la necesidad de volver a realizar investigaciones cuando los datos no son recuperables.

El segundo problema se refiere a la ausencia de estándares para la gestión y la preservación de los datos.

Finalmente, no existe integración entre bases de datos existentes, las cuales siguen normas propias o adoptan normas internacionales de manera independiente, generando dificultades para el intercambio de datos entre sí.

a) Falta de una plataforma de recogida de datos

Varios proyectos de investigación y mapeo de datos marinos de la plataforma continental de Brasil se han desarrollado en las últimas décadas, lo que genera una gran cantidad de información. Parte de esta información fue publicada en trabajos científicos, pero sólo una pequeña cantidad de ella fue incluida en las bases de datos que sirven como plataformas para la difusión del conocimiento. Según el informe de la ODE, y que bien sirve para representar el escenario global, se estima en un 58% el número de investigadores que desearían utilizar datos primarios ajenos, mientras que un 25% aproximadamente manifiesta problemas en compartir los suyos (Reilly, 2011). Conforme Bonetti (2012) citado por Castro (2012), en Brasil, debido a la falta de acceso a una plataforma estructurada, que proporciona datos de referencia con acceso abierto, muchos

científicos y estudiantes graduados terminan perdiendo gran parte de su tiempo de investigación en la generación de información básica, lo que limita el alcance de sus estudios. El autor añade que "muchos estudiantes de doctorado pasan mucho de su tiempo en la búsqueda de la estructura primaria de una base de datos. A menudo, este esfuerzo no sería necesario porque los datos ya existen, sin embargo, están dispersos, no son interoperables, o no están disponibles de forma abierta".

La falta de una infraestructura de datos oceanográficos disponibles de forma abierta se convirtió en un obstáculo para el avance de la investigación científica brasileña en diversas áreas. Esta escasez obliga a los investigadores a adoptar alternativas metodológicas para cruzar los datos principales, que podrían plantearse rápidamente en una plataforma para reunir datos de referencia financiados con inversión pública. Este escenario se explica en parte por la ausencia de esfuerzos coordinados entre universidades/centros de investigación, agencias gubernamentales y organizaciones no gubernamentales que tienen como objetivo, o praxis, la estructuración y normalización de los datos generados en estos estudios. Por lo tanto, la realidad actual muestra una pulverización de información desconectada, una pérdida sustancial de datos primarios y la imposibilidad de manipulación y acceso a la información. El daño científico, logístico y financiero de esta realidad es inmenso. El escenario brasileño, en este sentido, es antagónico a las tendencias internacionales que han alentado la organización de programas y la difusión de información como base para la investigación científica, en particular en lo que respecta a la difusión de datos de carácter ambiental. El desarrollo de estos programas no se limita a inventario, pero se refiere a la creación de una nueva "mentalidad" cuando se trata de la necesidad de compartir datos, sean brutos o catalogados.

Estas fallas estructurales aumenta la necesidad de mecanismos eficientes de divulgación científica, especialmente con respecto a la disociación de los datos válidos y útiles, de los contenidos innecesarios u obsoletos para investigadores, centros de investigación y universidades. Por lo tanto, es evidente que la organización adecuada de los datos en los ecosistemas marinos y polares[41] en

[41] En esa tesis, a efectos de clasificación del objeto de estudio global, los datos marinos y polares serán tratados como datos oceanográficos, así integrados como un único conjunto.

el caso brasileño implica la necesidad de revaluación del método de gestión en varios pasos, desde la adquisición hasta su archivo, control de calidad y su posterior difusión.

b) Ausencia de normalización

En Brasil, la gestión de datos oceanográficos se lleva a cabo por diversas instituciones y centros de investigación que utilizan unas reglas de normalización propias. Esta incompatibilidad de formatos de registros impide el intercambio sistemático de información, lo que dificulta establecer conexiones que permitan un diagnóstico amplio sobre los procesos de estudios del entorno oceanográfico.

La falta de estandarización de datos adecuado tiene varias razones, como la falta de conocimientos técnicos, la ausencia de recursos para la sistematización y la existencia de prácticas aún arraigadas que mantienen a los datos tan solo para uso interno en centros de investigación.

Cuando los investigadores se enfrentan a la necesidad de enfrentar las fuentes de datos oceanográficos y los intervalos de tiempo diferentes, encuentran una variedad de formatos y métodos de organización de los datos, centrado en soportes tecnológicos que impiden compartir (disquetes preservación, CD-ROM, archivos, etc.), los métodos de gestión específicos de los centros de investigación, además de no utilizar las normas establecidas y utilizadas a nivel internacional, causando barreras para participar en proyectos dentro y fuera del país; además de la duplicación de esfuerzos para ajustar los datos anteriores como base para futuras investigaciones.

c) Dispersión de bases de datos

En general, los bancos de datos de Brasil se han desarrollado en los últimos años por diferentes equipos que presentan fallas en la documentación y manipulan una serie de investigaciones que tienen contenidos y objetivos similares. De este modo se dispersan en diferentes repositorios de centros de investigación brasileños, como oficinas hidrografía, servicios geológicos, las

autoridades locales, agencias ambientales, institutos de investigación y universidades. Así pues, en Brasil, hasta el momento[42], no hay una base de datos estructurada en materia de investigación que integre los datos oceanográficos y el Programa Antártico Brasileño (Proantar)[43] que posibilite describir las características de la productividad brasileña en grandes áreas de los estudios marinos. Las bases de datos, tanto físicas, químicas, biológicas y geológicas, aún son insuficientes para establecer valores promedios confiables, por lo que es necesario seguir poblándolas con la recopilación de nueva y mejor información.

Existen diversas bases que se mantienen y son administradas por universidades, centros de investigación y agencias gubernamentales, lo que permite el libre acceso tanto a los datos en bruto y la información procesada. El OBIS (Ocean Biogeographic Information System) y el MGDA (Marine Geophysical Data Access) sirven como ejemplos más extendidos en las bases de datos espaciales a nivel internacional. Iniciativas brasileñas del mismo tipo también pueden ser citados como el Banco Nacional de Datos Oceanográficos (BNDO), el Banco de Datos de Datos Ambientales para la Industria Petrolera (Banpetro), el Sistema de Información Ambiental para el Programa Biota/ FAPESP (SinBiota) entre otros. La adopción de estas bases puede ser una solución eficiente para el rescate de datos a veces no disponibles. Sin embargo, la mayoría de estos programas no han proporcionado integración con diferentes bases de datos ambientales, evitando, por ejemplo, una correlación con bases de datos geológicos, biológicos e hidrodinámicas con el objetivo de la planificación y gestión de las áreas marinas protegidas, especialmente en las escalas locales y regionales.

42 El año de referencia 2013

43 Las actividades científicas de Proantar incluyen estudios e investigaciones en trece áreas de la ciencia.

5.9 Conclusiones

5.9.1 Resultados del estudio de usuarios

Las capacidades de la oceanografía operacional brasileña se han desarrollado al grado que, para investigaciones globales y regionales, varios centros producen rutinariamente datos sobre el estado actual del océano, incluyendo análisis y previsiones. Los científicos están de acuerdo que la mayor parte de la infraestructura operacional puede ser armonizada y compartida para el beneficio de todos, cambiando de independientes a centros multidisciplinares integrados; esto ha sido una de las lecciones clave de nuestra investigación.

Los investigadores entrevistados están de acuerdo que los datos oceanográficos son sumamente diversos, cubriendo una amplia gama de escalas espaciales y abarcando información capturada remotamente, mediciones *in situ*, y simulaciones numéricas. Según las respuestas obtenidas, para extraer la máxima cantidad de información de estas fuentes de datos ha sido necesario desarrollar y desplegar nuevas plataformas tecnológicas que permiten descubrir, compartir, visualizar y analizar esta información. El análisis de los investigadores y las respuestas de la encuesta demuestran una visión crítica del escenario brasileño en la cual apuntan que el futuro éxito de la oceanografía operacional depende de que la comunidad continúe trabajando junta para coincidir en estrategias en común, en particular:

- Formatos de archivos comunes y estándares de metadatos: La adopción del NetCdf ha sido muy exitosa para la armonización de tablas de datos tales como la salida de modelos numéricos y productos de análisis de satélite. La comunidad requiere ahora un estándar equivalente para datos in situ y de sensores remotos; sin tal estándar, será muy difícil construir los sistemas de datos y las aplicaciones requeridas por los usuarios de datos oceanográficos. Además, tiene que haber un acuerdo más amplio sobre como representar el uso de una estandarización integrada de metadatos y sobre la responsabilidad para un control efectivo de la área/alcance espacial y temporal de los conjuntos de los datos, y enlaces a información adicional como documentación.

- Componentes básicos en común para aplicaciones: La creación de una infraestructura de servicio de datos compartida sólo prosperará si servir al principal objetivo de ser de utilidad para los usuarios reales. Los investigadores utilizan a menudo herramientas para visualización y análisis de datos oceanográficos. Para estos usuarios, los portales actuales Web y herramientas de escritorio para propósitos generales, no poseen la funcionalidad específica que ellos necesitan. La comunidad oceanógrafa debe desarrollar componentes básicos de buena calidad y reutilizables, que puedan ser utilizados de varias maneras, es decir, para permitir desarrollar nuevas aplicaciones. Tales componentes básicos incluirán componentes de interfaz de usuario (como mapas interactivos) y servicios Web para acceder a datos y catálogos.

Existe la necesidad de desarrollar una plataforma única de gestión de datos, que esté estrechamente comprometida con los investigadores, para asegurarse de que los futuros sistemas de datos cumplan con sus necesidades. Hay un fuerte movimiento hacia la gestión de los repositorios de datos oceanográficos que atienda demandas específicas. Por ejemplo, en Europa el MyOcean distribuirá un sistema de datos pre-operacional que proporcionará paquetes de datos a los usuarios de datos oceanográficos de muchas disciplinas por todas Europa. Seria posible desarrollar sistema similar para atender las necesidades específicas de los investigadores y instituciones oceanográficas brasileñas.

5.9.2 Gestión de los datos

Un aspecto esencial de las acciones de articulación y interoperabilidad de datos se refiere a la necesidad de crear normas y estándares semánticos que establecen una base común para diferentes proyectos, además de reunir a los aspectos de diferentes disciplinas y campos del conocimiento bajo un mismo paraguas conceptual. En Brasil, el Decreto N ° 6666 de 27 noviembre 2008 estableció la Infraestructura Nacional de Datos Espaciales (INDE) y lo definió como el "conjunto integrado de tecnologías; políticas; y los mecanismos de coordinación y procedimientos de seguimiento; normas y acuerdos necesarios para facilitar y organizar la generación, almacenamiento, acceso, intercambio, difusión y uso de los datos geoespaciales de origen federal, estatal, del condado y municipales" (Brasil, 2008). En un país de dimensiones continentales como Brasil, con una gran falta de información adecuada para la toma de decisiones sobre cuestiones ambientales, compartir datos del oceano tiene un enorme potencial, aumentando el potencial de la gestión estratégica oceanográfica y que influye en la planificación, ejecución y seguimiento de las políticas públicas y el sector privado. Ejemplos internacionales de la implementación exitosa de la IED son notorios, especialmente para datos marinos. Se destacan en este contexto la INSPIRE (Infraestructura de información espacial en la

Comunidad Europea), que sirvió de base para el programa europeo de datos marinos EMODNET (European Marine Observation and Data). En Brasil, los esfuerzos a nivel nacional se han organizado desde el proceso de implementación del programa INDE/Brasil (Infraestructura Nacional de Datos Espaciales), coordinado por la Comisión Nacional de Cartografía (CONCAR), cuyo objetivo es la gestión de pedidos y el almacenamiento y el acceso de los datos geoespaciales y metadatos a nivel nacional (Brasil, 2008). Sin embargo, bajo el INDE, no existen directrices específicas relativas a los datos marinos o costeros. Es importante destacar que el desarrollo de una IDE no debe limitarse a los aspectos técnicos y operativos relacionados con las tecnologías de gestión de datos, sino también las acciones políticas y administrativas que proporcionan una amplia cobertura y difusión de su esencia y de la documentación (metadatos).

Los programas de interoperabilidad y de articulación - Además de la normalización y el establecimiento de la gestión y la difusión de los sistemas de datos (IDE), la implementación de programas de integración de datos marinos deben ir acompañadas de las redes locales del proyecto y de la organización y la cooperación de los grupos de trabajo especialmente en un país del tamaño de Brasil. El ICAN (International Coastal Atlas Network) es un ejemplo de red que comparte experiencias y estándares desde un enfoque multidisciplinario, cuyo principal objetivo es crear un medio para compartir datos e información marina y costera por la WEB. Otros ejemplos de la discusión y la regulación de los datos costeros son el Marine Metadata Interoperatibility Project, Network Marine Research Institutes and Documents - MARENET y el NOAA National Marine Data Center. Gran parte de estos programas se basa en la necesidad latente de desarrollar la investigación interdisciplinaria y multi-temporal para apoyar la comprensión de la dinámica de los ambientes marinos y costeros en un contexto de cambio ambiental global y así contribuir al desarrollo de estudios integrados a las acciones de conservación y para políticas públicas.

La UNESCO también ha desarrollado un papel clave en la organización de grupos de trabajo, a través de la Comisión Oceanográfica Internacional

(COI). Entre sus programas, el IODE (International Oceanographic Data and Information Exchange) subvenciona acciones de gestión y acceso a los datos marinos que prestan apoyo técnico, educativo y financiero para establecer bases sólidas para el desarrollo de un sistema mundial de información de datos marinos y costeros en menos de una década. La Marina de Brasil, a través de la Dirección de Hidrografía y Navegación (DHN) es la institución nacional cuyas funciones son promover y coordinar la participación del país en las actividades de la COI relacionadas con los programas de Servicios Oceánica, y servir Banco nacional de datos Oceanográficos (BNDO) y el Centro Depository COI, integrando así el Sistema Global para datos oceanográficos. Entre los programas de Servicios Oceánicos de la COI de hoy en día, se está haciendo hincapié en el desarrollo e implementación del Sistema Mundial de Observación de los Océanos (GOOS), cuya supervisión está a cargo de la Comisión Interministerial para los Recursos del Mar (CIRM).

En el caso brasileño, los datos son procesados e interpretados de manera interdependiente y mal integrados tanto geográficamente (no articulados) en relación a su disciplina y especificidad. Tal enfoque es generalmente apoyado por la investigación temática que tiene poco o ningún medio de gestionar y compartir información. Fundamentalmente toda la investigación marina y costera en Brasil hasta principios del siglo XXI, se basó en una colección de valoración de recogida de una gran cantidad de datos, dirigido a la caracterización de un aspecto particular de la investigación (por ejemplo, la taxonomía, ecología, geoquímica , geología) y el mantenimiento de estos datos cautivos para tales usos (Conti et al., 2013). Por lo tanto, un aspecto prioritario que impregna toda la discusión sobre los nuevos proyectos de gestión de datos se refiere a la necesidad del levantamiento y la catalogación de datos asociados a los proyectos ya ejecutados, ya que una pequeña parte se encuentra estandarizada y disponible en las publicaciones de acceso abierto (Conti et. al. 2013).

Además, la integración de los datos dentro de una misma disciplina, o en proyectos similares basados en colecciones o recopilación de datos en un contexto multidisciplinar, incluso sin tener análisis integradas en sentido

estricto. Proyectos como el "Atlas geográfico de las Zonas de Brasil costeros y oceánicos" (IBGE, 2011) incluyen la integración de bases de datos cartográficas de diferentes naturalezas. Todavía, son por lo general de uso y acceso restringido a las instituciones generadoras y mantenedoras que mantienen y que se liberan por lo general sólo los productos finales (es decir, mapas , informes, artículos técnicos).

En el caso de los datos multidisciplinares y multiespaciais (articulados), informaciones completamente diferentes pueden se correlacionar y ser disponibilizadas espacialmente, como ocurre en la mayoría de los inventarios de datos espaciales basados en geoportales. Actualmente, la mayoría de las discusiones sobre los aspectos técnicos y prácticos de las bases de datos espaciales marinos y costeros si encuentran en este nivel (Conti et. al. 2013).

En un nivel superior, se incluyen programas totalmente integrados, donde no hay solo una regulación y organización de los datos, sino también el desarrollo efectivo de herramientas de geoprocesamiento combinadas que producen necesariamente nueva información derivada de los datos originales. Un ejemplo de estos proyectos de carácter en las zonas marinas es la Red de Proyectos de Biodiversidad Marina (BIOMAR), que reúne proyectos patrocinados por el Programa Petrobras Socioambiental. Sus productos van más allá de la cartografía y el suministro de datos para la biodiversidad destinadas a la producción de nueva información ambiental. Se logra la plena utilización de los datos sólo en el nivel más alto, ya que la información de diversas fuentes y análisis de escalas pueden ser integrados y procesados, con el fin de generar modelos predictivos complejos y altamente estructurados. Pocos ejemplos en el mundo tienen este nivel de integración, pero estas excepciones exitosas han mostrado prometedores con un intento de comprender la naturaleza extremadamente diversa del medio marino y la interconexión de los procesos activos. Estos niveles dependen invariablemente a la entrada de datos de acciones de colaboración y requieren una gestión activa, no sólo para organizar y estandarizar las informaciones (y necesariamente precedido por una base de IDEs), sino también para diseñar e implementar nuevas herramientas

analíticas. Las organizaciones gubernamentales y no gubernamentales que se ocupan de la gestión de datos están invirtiendo en el desarrollo de diversas herramientas para compartir en el nivel de análisis para el futuro.

6 PROPUESTA DE MODELO

6.1 Introducción

Recogiendo el análisis del trabajo llevado a cabo con el estudio de usuarios, análisis de la situación internacional y en Brasil, las prácticas y las estrategias de colaboración en el escenario internacional identificadas como parámetro en el ámbito de la gestión de datos oceanográficos, a continuación se presenta una propuesta de los elementos que debe contener el modelo global de gestión entre los diferentes agentes involucrados (gobierno, comunidad científica e investigadores), que permita coordinar, ampliar y canalizar la colaboración e interoperabilidad del entramado de centros de investigación, repositorios y proyectos que actúan en la recogida y preservación de datos marinos de manera directa o indirectamente en Brasil.

Este modelo deberá concretar un conjunto de acciones diversas con el objetivo de sensibilizar los gestores de los datos oceanográficos, ampliar redes de colaboración, canalizar ayudas, generar espacios y mecanismos efectivos que atentan para los estándares internacionales de normalización. Se debe tener en cuenta que algunas de las acciones específicas que se propongan en el modelo, se deben ejecutarlas, a modo de validarlas, en cada centro de investigación y área específica. El modelo pondrá especial énfasis en la transferibilidad del mismo, de modo que pueda ser aplicado en diversos sectores sin que las diferencias institucionales y económicas constituyan un impedimento a su aplicación.

La propuesta de modelo definida contiene una metodología concreta para poder coordinar, ampliar y canalizar la gestión de los datos oceanográficos, y que además contemple el desarrollo de acciones específicas para su implantación. Se trata de un modelo que se da a conocer como un instrumento global con diferentes líneas de actuación para trabajar, desde la definición de parámetros para inclusión, manejo y consulta de los datos. El objetivo es intentar agrupar en la medida de lo posible una gran parte de conceptos estructurales para el desarrollo de una plataforma de datos de investigación oceanográfica, a través de este mecanismo. La filosofía de

fondo es promover y dar soporte al trabajo en red para incrementar una infraestructura nacional que atienda los requisitos y parámetros para la internacionalización de la ciencia marina brasileña.

El análisis del estudio de usuarios y de la situación actual de la gestión de datos verificada a lo largo de nuestro estudio muestra que el desarrollo de una infraestructura de datos científicos oceanográficos para Brasil debe ser evaluada desde la perspectiva del fortalecimiento del modelo vigente. Es necesario establecer una base para el fomento de prácticas colaborativas que mejoren el trabajo científico y permitan el aumento del uso y explotación de los datos de investigación y por consecuencia faciliten su reuso. Para eso, hemos construido un modelo de datos como parte del sistema de gestión de datos brasileño, combinando una especificación oceanográfica (Arc Marine) y un estándar internacional de metadatos (la ISO 19115) en un sistema de gestión de datos de código abierto (PostgreSQL) que permite accesorios (agregados) GIS (PostGIS). Arc Marine proporciona los componentes oceanográficos específicos del modelo de datos utilizando un juego de entidades y atributos bien definidos.

Finalmente el modelo deberá servir para:

a) Sensibilizar al conjunto de involucrados con la gestión de datos oceanográficos acerca de la necesidad de colaborar en el desarrollo de estrategias y métodos de colaboración para el desarrollo de una infraestructura integrada de datos marinos.

b) Desarrollar mecanismos que permitan una clara identificación de los principales problemas que se deben llevar a cabo y que son susceptibles de cambios (identificación de necesidades/problemas, desarrollar repositorios de datos integrados, etc.).

c) Concretar qué aspectos estructurales son necesarios crear, modificar etc. para poder coordinar y canalizar las diferentes acciones que integran el modelo (necesidad de revisar la adopción de estándares, etc.)

6.2 El modelo Arc Marine

Hemos construido un modelo de datos como parte del sistema de gestión de datos brasileño, combinando una especificación oceanográfica (Arc Marine) y un estándar internacional de metadatos (la ISO 19115) en un sistema de gestión de datos de código abierto (PostgreSQL) que permite accesorios (agregados) GIS (PostGIS). Arc Marine proporciona los componentes oceanográficos específicos del modelo de datos utilizando un juego de entidades y atributos bien definidos.

Conforme demostraremos, los registros de metadatos del modelo Arc Marine reúne la posibilidad de indexar datos e investigaciones realizadas en las diferentes disciplinas con información para su difusión internacional, y lo que es más importante, su intercambio. Para este propósito, se hace necesaria la presencia de un centro distribuidor de metadatos, con características implícitas tomadas desde dos puntos de vista: 1) desde una base de datos distribuida que tiene una copia completa de la base de datos, o partes del mismo, en más de un lugar, y 2) desde una base de datos replicada que mantenga copias completas de la base de datos en varios lugares, principalmente para reducir los problemas en el caso de fallo de la base de datos centralizada. La base de datos particionada está dividida de modo que cada sitio tiene una copia completa de la base de datos. Este tipo de base de datos proporciona una buena velocidad de respuesta a los archivos localizadas, no hay necesidad de replicar todos los cambios en múltiples ubicaciones.

En el planteamiento de la infraestructura de datos resulta primordial el diseño de una estructura de datos común a todos los ámbitos de la investigación oceanográfica, definida por unas normas o estándares, en algunos casos, y en otros por diseños o protocolos desarrollados por las partes implicadas. Las estrategias de intercambio de un sistema de base de datos distribuida

(BDD) consiste en una lista de nodos, cada uno de los cuales puede participar en la ejecución de las transacciones que accedan a datos en uno o más nodos. En un sistema de base de datos distribuida, la base de datos es almacenada em varios equipos (nodos). La principal diferencia entre la base de datos centralizada y los sistemas distribuidos es que los primeros datos se encuentra en un lugar, mientras que en la otra los datos residen en múltiples lugares. Esta distribución de los datos es el objeto de muchas preocupaciones y dificultades.

Los procesadores en un sistema distribuido pueden variar en tamaño y función, y pueden incluir computadoras personales, estaciones de trabajo, sistemas grandes y el uso de sistemas informáticos en general. Estos procesadores son generalmente llamados nodos, dependiendo del contexto en el que se mencionan. utiliza principalmente el término nodo (lugar, posición) con el fin de hacer hincapié en la distribución física de estos sistemas.

Para ello se tendrán en cuenta las siguientes lineas generales de la propuesta para un modelo de gestión de datos oceanográficos en Brasil:

a) Organización de datos y desarrollo de un catálogo de metadatos

Las respuestas obtenidas en el estudio de usuarios evidencian que la gestión de metadatos pone exigencias especiales a los administradores, lo que hace que las bases de datos tradicionales sean incapaces de resolver fácilmente. Un aspecto muy importante para el desarrollo del uso de bases de datos para diferentes grupos de usuarios con intereses múltiples, es la documentación de su contenido. Sin la documentación adecuada hace que sea difícil para los usuarios encontrar los datos que necesita para sus aplicaciones y entender su significado. Una vez encontrados los datos, por lo general es necesario conocer la forma en que fueron recogidos y que precisiones tienen.

Algunas de las dificultades inherentes a la fase de organización de datos están asociados con la determinación de parámetros de metadatos, o la

calidad de los datos, tales como índices de exactitud de posicionamiento, la descripción de los métodos de recolección, almacenamiento, tratamiento en el laboratorio, y la integridad de los datos. Dicha información, cuando existe, conforme demuestran los resultados de la encuesta (cap. 6) además de catalogadas, también deben estructurarse de acuerdo a los estándares internacionales de metadatos a través de un conjunto de información esencial para ayudar en la localización, descripción y comprensión de los datos geoespaciales, es decir, la ejecución de un proyecto de Infraestructura de Datos Espaciales (IDE). Estas normas cubren el contenido y la semántica de los metadatos, incluyendo su documentación detallada y también su representación digital. El propósito para el establecimiento de estas normas es proporcionar terminología y definiciones comunes para los conceptos relacionados con estos metadatos geográficos.

La estructuración de un catálogo de metadatos tiene como consecuencia directa, la identificación de lagunas en el conocimiento, tanto en términos espaciales como en términos conceptuales. Mediante la creación de normas, los metadatos especifican el contenido de la información a un conjunto de datos geográficos digitales, pues catalogan la Información Geográfica (IG), describen sus características, las condiciones, la calidad, etc. Como hemos visto, los estándares han sido y están siendo propuestos por consorcios de bases de datos internacionales y han llegado beneficios tangibles con su adopción. En este contexto, la formulación de modelos de interoperabilidad es esencial para el registro de la IG en las IDEs, lo que posibilita analizar el comportamiento de los sistemas desde distintos enfoques o niveles (Machado; Bermejo, 2010).

El modelo de interoperabilidad propuesto para las IDE estudia los niveles definidos en los modelos aplicados a los sistemas de sistemas, es decir, los requisitos existentes en sistemas de gestión de datos existentes en Brasil. En el caso brasileño, los modelos de interoperabilidad basados en los metadatos requieren una metodología original que permita analizar la interoperabilidad proporcionada por los mismos y que facilite la creación automática de metadatos y de una metodología. El método propuesto para crear automáticamente metadatos estructura el proceso de compilación y

tratamiento de la información, compone y almacena el metadato estandarizadamente y puede integrar-se en los flujos de trabajo de la IG.

El análisis de la interoperabilidad que proporcionan los metadatos de la norma ISO 19115 define el modelo conceptual requerido para describir la información y servicios geográficos. La necesidad de integrar los metadatos que demuestra la encuesta a expertos (cap. 6) ha disipado la incertidumbre en torno a la subjetividad de la identificación de la interoperabilidad proporcionada por los metadatos.

b) Bases de datos y estrategias de intercambio

Tal como se ha visto anteriormente, el desarrollo de bases de datos de red distribuida georreferenciada puede seguir dos formas básicas de la estructura: lo que se llama "bases de datos Localizados" (BDL), donde el conjunto de información, datos y metadatos se establece físicamente en un centro (servidor) y la "Base de datos Distribuidores", en el que la información se pulveriza en varios BDLs y es agregada en un nodo que la codifica y integra.

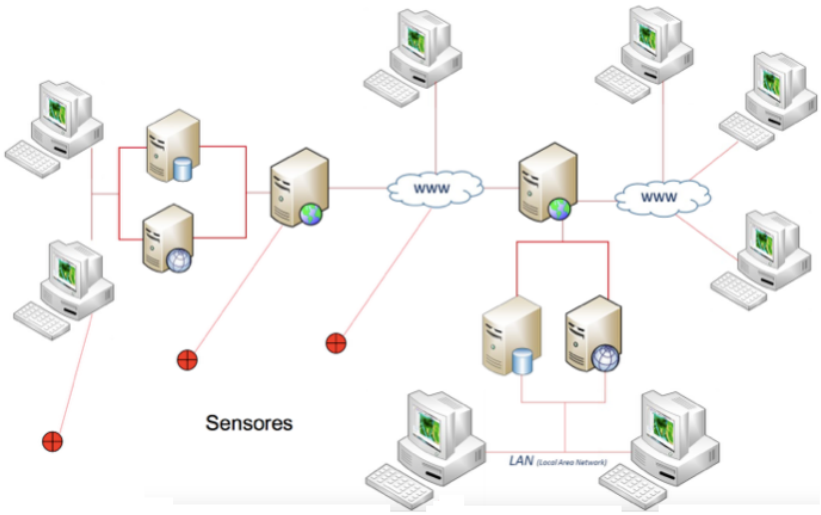

Figura 24: Bases de datos Localizadas y Base de datos Distribuidoras

La encuesta con los investigadores evidencia que gran parte de los sistemas desarrollados en Brasil en los últimos años tiene un carácter híbrido, con un

poco de información localizada "in house" en los servidores y añadiendo al mismo tiempo información como un nodo. Independientemente de la ubicación local o distribuida, los sistemas de bases de datos georreferenciados compartidos deben tener necesariamente un sistema con semántica y ontologías que permite al usuario acceder, manipular y analizar información geográfica dentro de las normas y estándares establecidos. Por ejemplo, para buscar un término "línea de costa" existe la posibilidad del retorno de diversas informaciones con niveles de precisión y formatos, como archivo vectoriales derivados del sistema GEBCO (General Bathymetric Charts of The Oceans), líneas escaneadas a partir de mapas topográficos, líneas de referencia (como la marina o el Instituto Brasileño de Geografia y Estadística). Por otra parte, el término "línea de costa" debe organizarse con el fin de estar asociado con referencias a la "línea costera", "costa", como los términos en Inglés “shoreline” e “coastline”.

Mediante el establecimiento de una base de datos del tipo BDD o Híbrido, también es necesario contar con un patrón claro de interoperabilidad con los sistemas de alimentación de la BDL. En el caso de un sistema que implica la biodiversidad marina, es imprescindible que exista una coordinación con los sistemas internacionales, tales como el World Register of Marine Species (WoRMS) y el OBIS, y nacionales, como *el* Sistema de Informação Ambiental para o Programa Biota/FAPESP (SinBiota).

c) Desarrollo de un portal para la difusión y distribución de la información

A partir de una infraestructura de datos y metadatos, que implica tecnologías, políticas y reglas establecidas, la distribución de la información georeferenciada de Internet de un portal de información geográficas, ha sido la base para soluciones de intercambio y difusión de información.

Tal como vimos en el cap.5, en referencia a los repositorios internacionales, los Estados Unidos son una referencia en relación a los servicios de búsqueda en todos los tipos de datos geográficos producido por el sector público con el fin de facilitar su uso y distribución. El proyecto del portal

estadounidense llamado Geospatial One-Stop Portal se basa en inversiones ya realizadas por el gobierno federal en la constitución de una Infraestructura Nacional de Datos Espaciales (INDE) y trata de hacer que el acceso sea fácil, sencillo y barato para todos los niveles de los datos del gobierno e información geográficas[44].

La política de difusión de datos geográficos sigue el principio de libertad de saber qué y cómo los datos geográficos son accesibles. Por lo tanto, un portal de información geográfica, incluyendo los datos oceanográficos, proporciona, si no los propios datos, al menos los metadatos para evaluar su idoneidad para el uso previsto, incluyendo la exactitud y precisión, así como la información acerca de las políticas de acceso practicados por sus productores.

Una serie de proyectos de atlas costeros y marinos se han desarrollado dentro de este contexto, así como las acciones supranacionales, con un amplio apoyo de las agencias ambientales nacionales y locales, que implica la integración de los datos bióticos y abióticos como base para el desarrollo de políticas ambientales comunes. Tales esfuerzos se centran principalmente en el desarrollo de métodos y estrategias que permitan la integración de una gran cantidad de información de la naturaleza, alcance, precisión y estructura diferente. Ejemplos como los proyectos Marine Irish Coastal Atlas (MIDA)[45], Belgian Coastal Atlas (De Kustatlas)[46], Venice Coastal Atlas[47], Oregon Coastal Atlas[48] son cada vez más comunes, y en muchos casos, sirven para apoyar el desarrollo de programas amplios y diversos, tales como OCEANGIS NOAA (EE.UU)[49].

La creación de un Web-Atlas (y la infraestructura y la organización de las bases de datos que le precede) permite no sólo los datos de entrega y la

---

[44] http://www.geodata.gov

[45] http://mida.ucc.ie/contents.htm

[46] http://www.kustatlas.be

[47] http://atlante.silvenezia.it

[48] http://www.coastalatlas.net

[49] http://www.pmel.noaa.gov/vrl/OceanGIS

información, sino también el análisis y procesamiento. Aunque gran parte de los sistemas actuales todavía presentan un carácter embrionario en este sentido, el consenso dentro de la comunidad científica es que el suministro de una cantidad cada vez mayor de datos junto con el desarrollo de los procesos informáticos de red (cloud computing) se mueve de los procesos de análisis y manipulación de datos de la red en sí misma y, en consecuencia, los sistemas web-Atlas pueden servir como base para proyectos de investigación dirigidos a derivados del análisis y aplicación de datos, especialmente con respecto a la gestión costera y los programas de conservación, tales como el Programa Nacional de Gerenciamento Costeiro (GERCO)[50] o Programa de Planificación Espacial Marina y Costera (PPEMC).

La acreditación del Banco Nacional de Dados Oceanográficos (BNDO) como un Centro Nacional de Datos Oceanográficos (NODC), de conformidad con la nueva estrategia Oceanográfica de datos y la información Gestión de la propiedad, aprobada por la Comisión Oceanográfica Intergubernamental (COI) para el Programa para el Cambio Internacional de Datos y Informaciones Oceanográficas (IODE), es una herramienta estratégica de difusión de los datos científicos de la investigación brasileña que debe ser explorada con mayor detalle.

6.3 Diseño de la geodatabase

6.3.1 Concepto básico

Para la propuesta del modelo de la geodatabase, se buscó una referencia estándar que atienda las principales necesidades destacadas en el estudio de usuarios, pero no existe una acreditación universal para este tipo de almacenamiento de datos. Se encuentra, sin embargo, un modelo propuesto por la comunidad Arc Marine[51], este modelo estándar incluye las consideraciones necesarias para estructurar actividades de investigación marinas, ha sido creado por un conjunto de investigadores de prestigiosas universidades, cumple con tener una amplia cobertura de aplicación, que

[50] http://www2.mma.gov.br/sitio/index.php?ido=conteudo.monta&idEstrutura=26

[51] http.//dusk.geo orst.edu/djl/arcgis/

permite la implementación de herramientas desarrolladas para el trabajo en bases similares generados en grupos de investigación de diversas partes del mundo, al optimizar la integración e intercambio de datos (TAPIA, et al, 2006).

La *geodatabase* es el modelo de datos primario de ArcGIS. El nombre combina la palabra *geo* (como referencia a lo espacial) con *database,* específicamente un sistema relacional de base de datos (RDBMS). El término promueve la idea de que todos los datos SIG sean almacenados en una ubicación central para un fácil acceso y administración. Dado que todos los datos de una *Geodatabase* son almacenados directamente en sistemas gestores de bases de datos comerciales (Microsoft Access para *Geodatabase* personal y Oracle, IBM DB2, SQL Server o Informix para *Geodatabase* corporativa) o en sistemas de ficheros, éstos constituyen un repositorio común, único y centralizado para todos los datos geográficos de una organización.

La geotabase propuesta está concebida como un sistema de gestión de datos marinos organizado para compartir geometrías, con el objetivo de integrar la mayor diversidad de tipos de datos marinos posible. El sistema deberá hacer posible el registro, validación, búsqueda, recuperación, visualización, análisis y exportación de los datos, posibilitando la interoperabilidad con otros repositorios o nodos con mayor rango. En el Figura 7 se representa el concepto básico de esta idea aplicada al escenario brasileño. El esquema general de la arquitectura de todo el sistema se muestra en el apartado siguiente.

Figura 25: propuesta de esquema simples para un modelo de gestión datos oceanográficos en Brasil

Una vez disponibles los datos en el repositorio de datos, la geodatabase procurará añadir módulos analíticos y servicios concretos que potencien aún más la utilidad de los datos para la sociedad (navegación, pesquerías, investigación, uso recreativo del mar, aspectos sanitarios, etc.) haciendo más rentable el mantenimiento de un repositorio de datos marinos abierto a múltiples usos y usuarios.

6.3.2 Las especificaciones del Arc Marine

Las dimensiones territoriales de Brasil requieren un modelo integrado de gestión de datos compatible con el volumen de datos que genera. Para eso necesita organizarse en función del tipo de dato y no en función de aplicaciones concretas con fines específicos (gestión costera, navegación, etc.) como es habitual en el diseño de la mayoría de los sistemas de información geográfica.

La proposición del modelo está orientada al almacenamiento de forma integrada de cualquier tipo de dato marino, sea oceánico o costero, desde que tenga un referente geográfico, ampliando así sus posibilidades de explotación. Para esta finalidad buscamos un modelo de datos que cuente con un conjunto de entidades geométricas (puntos, líneas y polígonos) unificado. La opción adoptada fue el modelo *Arc Marine Common Marine Data Types* (Figura 8) desarrolladas por Wright y colaboradores (2007) debido a su estructura universal y flexible en relación a las posibilidades de gestión de datos oceanográficos.

El Arc Marine es un modelo bastante consolidado entre la comunidade oceanográfica. Como ejemplo, existen organizaciones como el Repositorio de datos marinos integrados de Canarias (REDMIC), un sistema permanente de almacenamiento sistemático, custodia y servicio de datos marinos. Se ha diseñado para Canarias, aunque haya sido concebido como proyecto piloto con el objetivo der ser replicado en otras regiones. El REDMIC posibilita que los datos marinos, de cualquier tipo (oceanográficos, tráfico marítimo, biodiversidad, etc.), sean incorporados de modo integrado en un mismo

sistema de información geográfica, "con lo que el esfuerzo de ponerlos en común se hace una sola vez al principio, para que luego pueden ser usados y combinados cuantas veces se desee con la máxima agilidad" (Machado, 2010). El modelo lógico de datos, desarrollado a partir de Arc Marine, es lo que permite la integración de los datos.

El US Geological Survey (USGS) y la Oficina de Massachusetts de gestión de zonas costeras presentan mapas de alta resolución del fondo marino. Las imágenes son exportadas en formato TIFF georreferenciado para su posterior análisis en el ArcMarine y se almacenan todos los datos espaciales dentro de una geodatabase basado en el modelo de datos ArcMarine (USGS, 2016).

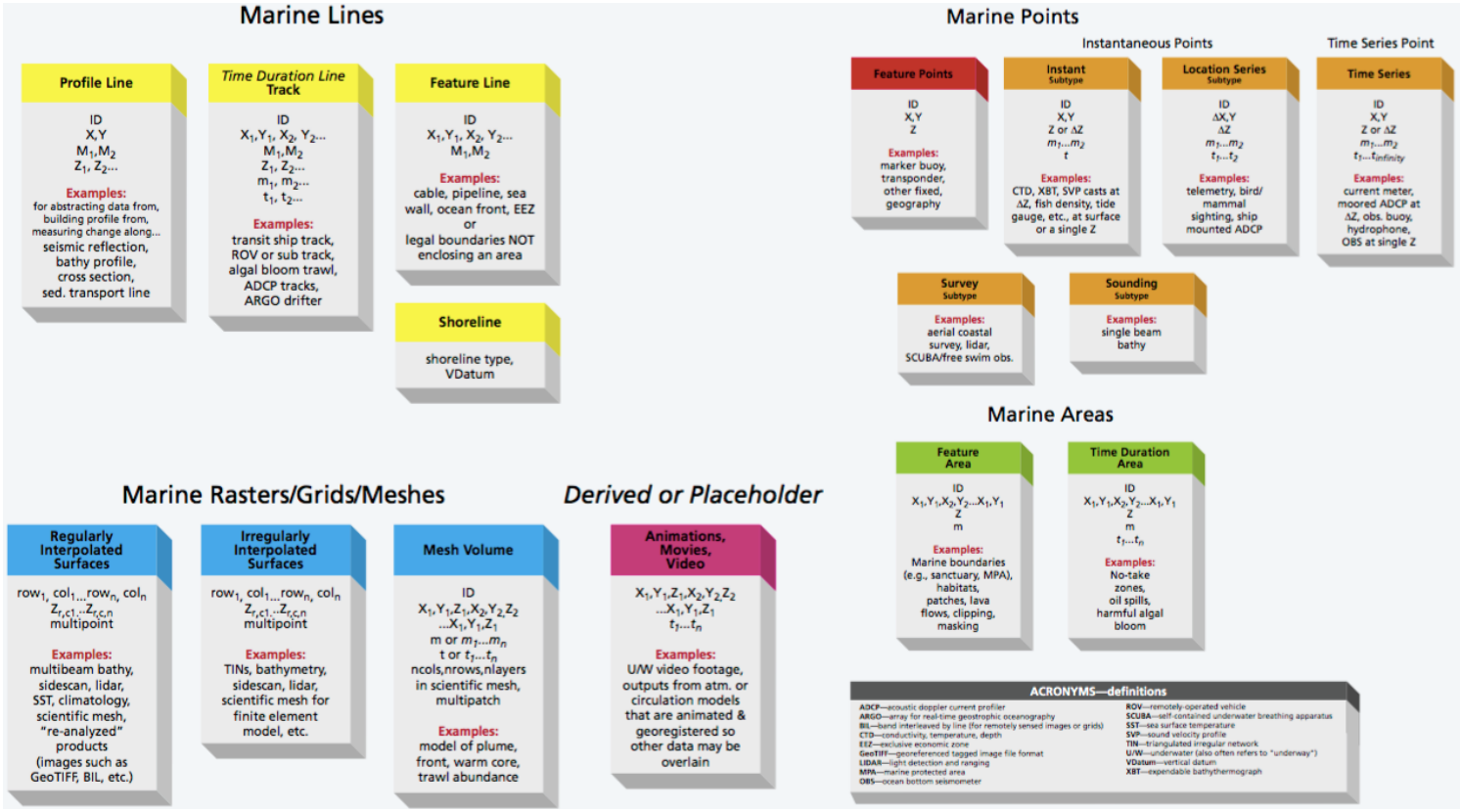

Figura 26: Diagrama de los tipos comunes de datos marinos Arc Marine (Wright et al. 2007)
Leyenda: **XY** son las coordenadas geográficas de latitud/longitud y **Z** la profundidad **m** representa el valor de lo medido, y **M** es la resolución de dicho valor
t representa el tiempo

En el Arc Marine los datos quedan vinculados al factor geográfico vía el sistema de geometrías compartidas, y se relacionan obligatoriamente con la actividad que los genera y toda la información asociada (metadatos). Para evitar la redundancia de información, ésta se estructura básicamente según la clásica secuencia de preguntas: quién, dónde, cuándo, qué y cómo, pudiendo ser compartida entre las distintas actividades. Mediante el establecimiento de un conjunto común de terminología de metadatos, definiciones y procedimientos de extensión, este estándar promove el uso adecuado y la recuperación de los datos geográficos. En este contexto, también se aprovechó el enfoque organizativo desarrollado por el Marine Institute de Irlanda[52] para su repositorio de datos marinos.

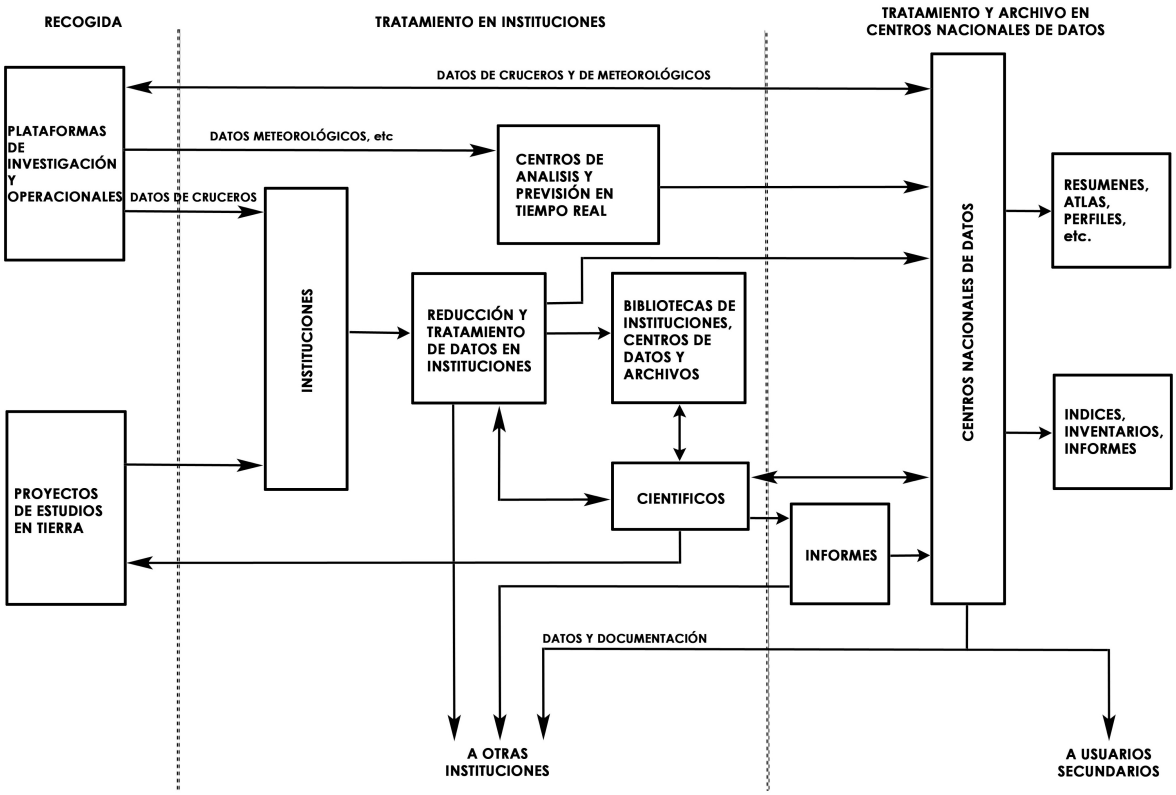

Figura 27: Guia para establecer un centro nacional de datos
Fuente: elaboración propia, con base en el el enfoque organizativo desarrollado por el Marine Institute de Irlanda para su repositorio de datos marinos.

[52] www.marine.ie

La especificación de Arc Marine es un esfuerzo que se enfoca en especificaciones abiertas suficientemente flexibles para permitir la separación lógica de juegos de datos internos y externos desde una perspectiva de gestión.

Para desarrollar una infraestructura de gestión de datos oceanográficos en Brasil, presentamos cuatro etapas del diseño de un sistema que atienda las necesidades para gestión de datos en ámbito nacional, que son:

-Diseñar e implementar la capa de datos e información del sistema de gestión de datos aprovechando estándares abiertos (p.ej. modelos de datos, estándares de contenido) al alcance viable para apoyar operaciones en tierra y en mar. Esto incluye aprovechar estándares abiertos para maximizar la extensibilidad del sistema.

-Probar las estructuras de datos incorporando juegos de datos existentes en la base de datos construida desde los modelos de datos y proporcionar estos datos a la comunidad local de usuarios. Esto incluye la incorporación de los metadatos pertinentes a los procedimientos y sensores utilizados para recoger los datos.

-Permitir la portabilidad de información en el sistema. El aspecto de la portabilidad permitirá que una parte de los datos relevantes al área local de operación sea movida a un entorno informáticamente aislado en el mar (p. ej. sin conectividad a Internet), mientras no incluyan datos de las áreas marítimas distantes, que son irrelevantes a las actividades en mar.

-Gestionar solo los datos de los cuales el laboratorio es responsable. Un problema para la conservación de los datos de investigación se encuentra en la preservación de la información contenida en bases de datos. La información solicitada por un investigador puede incluso existir en un determinado conjunto de datos, pero no tendrá ningún valor para ese investigador, si no tiene una manera de cuestionar su existencia. Por otra parte, hay casos en los que se permite el acceso a un determinado conjunto de datos, pero la lentitud del proceso de interpretar y extraer información contenida en el mismo, hace que su contenido sea menos valioso. Ambos problemas son bastante críticos

para la correcta conservación a largo plazo de los datos generados en la investigación científica, y como tal deben ser abordados en cualquier proceso curatorial de esta zona.

El modelo de información presentado aquí junta tres componentes clave en el sistema: el marco de Arc Marine; la Organización Internacional para la Estandarización (ISO) de Información-Metadatos Geográficos estándar 19115; y el sistema de gestión de base de datos de código abierto (DBMS) PostgreSQL con extensiones de Sistema de Información Geográfica (GIS).

La proposición del modelo esta estructurado de la siguiente manera: La sección 5.6 proporciona una introducción a los componentes principales incluyendo el marco de Arc Marine, el estándar ISO 19115 y el PostgreSQL DBMS.

Las secciones 6.3.8 y 6.3.9 describen las estructuras de datos que apoyan modelos de datos de punto y numéricos. La sección 6.3.10 introduce los componentes de linaje del sistema y el empleo de ISO 19115 para el rastreo de linaje, partes responsables, citas, y extensión geoespacial. La sección 6.3.11 describe el concepto de salida de usuario, una aplicación utilizada para acceder a juegos de datos externos, apoyados por otras organizaciones, mientras la Sección 6.3.12 presenta un análisis general del modelo.

6.3.3 Componentes

Esta sección introduce brevemente al modelado de datos, seguido de descripciones de los principales componentes utilizados en su diseño.

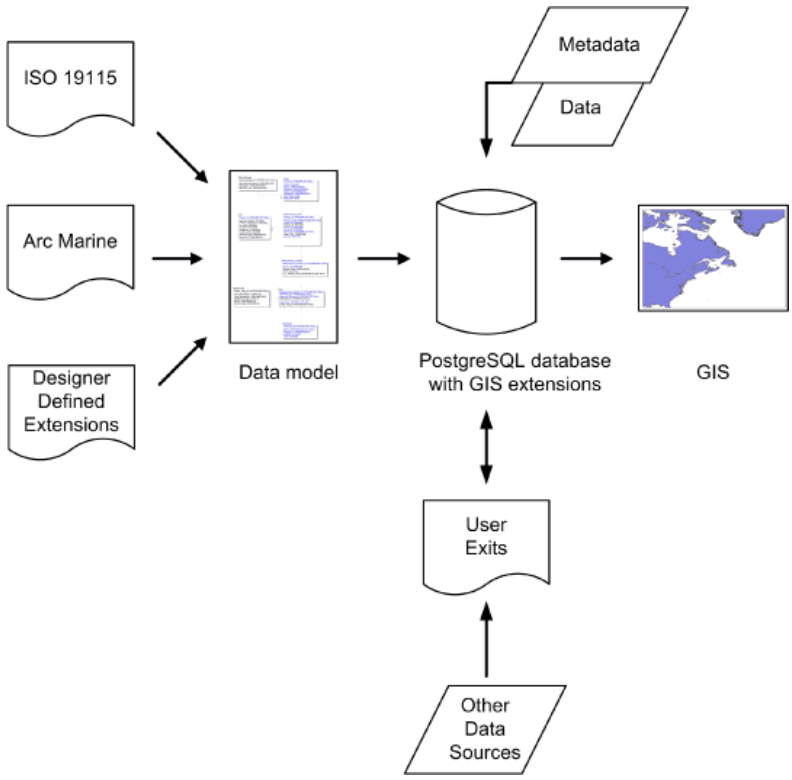

Figura 28: Ilustración de componentes que conducen a la base de datos final en el PostgreSQL DBMS
Fuente: Wright et al. (2007)

6.3.4 Marco de Arc Marine

Un marco es una estructura subyacente o un concepto alrededor del cual algo es construido. El marco de Arc Marine (Wright et al. 2007) identifica las estructuras de datos relevantes a la comunidad marítima en general. Estas estructuras de datos entonces pueden ser usadas como el corazón de un sistema de datos para manejar diversos datos oceanográficos, como los requeridos por los sistemas de gestión de datos en general.

El grupo de desarrollo de Arc Marine estaba formado por oceanógrafos físicos, geólogos marítimos y biólogos. El trabajo inicial mostró como el marco de Arc Marine podría ser usado para crear un modelo de datos (Lord-Castillo et al. 2009, Wright et al. 2007), con la implementación del modelo de datos en la línea de productos de Esri®GIS (Esri 2010). Sin embargo, muchos de los conceptos y técnicas de diseño del marco son independientes del proveedor (abastecedor). Esto quiere decir que las estructuras de datos que constituyen

el marco puede ser aplicadas a otros entornos de desarrollo. A su turno, esto quiere decir que Arc Marine puede ser considerado un marco abierto, aplicable a muchos desarrollos y escenarios de aplicación.

Como un marco, el Arc Marine no es un simple modelo de datos, sino más bien una estructura conceptual de datos para desarrollar un modelo de datos de uso específico u orientado a una aplicación. El marco de Arc Marine incluye estructuras de datos que apoyan la gestión de medidas oceanográficas puntuales, series de tiempo, líneas de perfil, etc. Usando estos objetos generalizados, entidades de modelos de datos pueden ser construidas para el área de aplicación específica. También, Arc Marine define un proceso de pensamiento formal para ser usado en la definición de cualquier entidad adicional.

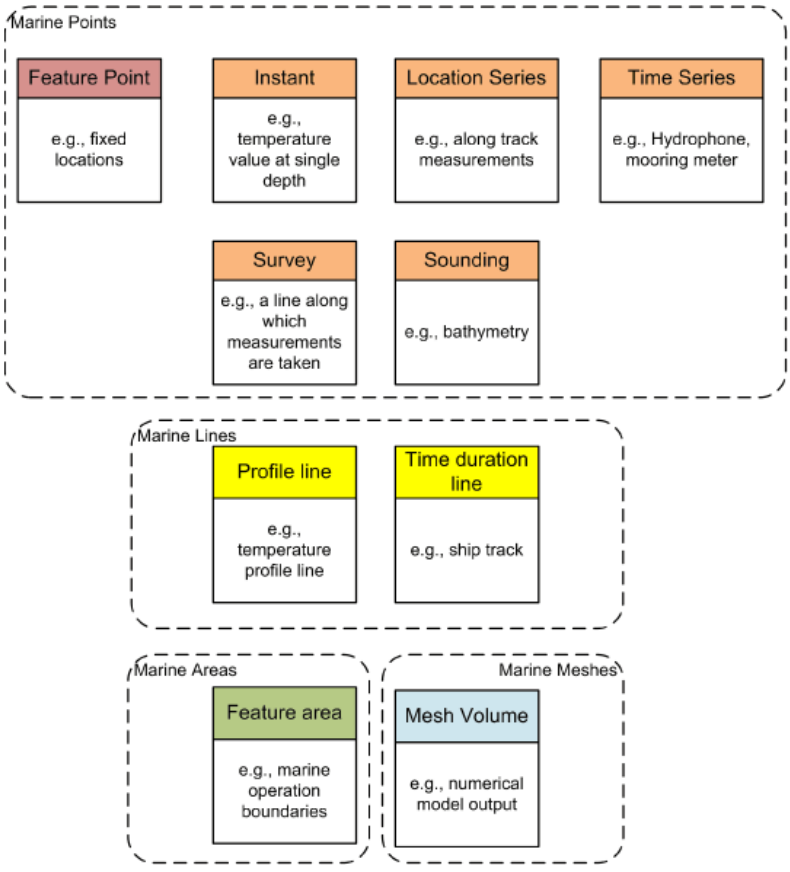

Figura 29: Subconjunto del marco de Arc Marine presentado en Wright et al. (2007)

Hay numerosas ventajas en la utilización de un marco abierto como Arc Marine. El marco abierto permite la opción de compartir el contenido de datos a través de otras aplicaciones que también implementen el marco, así como

también el potencial de compartir aplicaciones de software que utilizan datos dentro del sistema. La utilización también fomenta a extender el marco, como es el caso de la presente propuesta de un modelo de gestión de datos oceanográficos para Brasil.

Arc Marine proporciona el marco para el almacenaje de dos categorías primarias de datos: datos concretos (puntuales) y de red (malla). Las categorías para líneas y áreas geoespaciales están también presentes en el modelo de datos, pero estas categorías están apoyadas por los datos puntuales y de red del entorno.

Los datos de red son una forma particular de datos puntuales. Una red es una serie de puntos en una orden definido y predecible que representa una superficie interpolada regular o irregularmente (Wright et al. 2007). Los juegos de datos que han sido proyectados en una rejilla constituyen una forma de datos de red.

Las salidas (de datos) de un modelo de océano numérico hidrodinámico, también serían un ejemplo de datos de red. Sin embargo, en el modelo de datos propuesto, la inclusión de salidas oceánicas de modelo numérico, son un caso atípico. Esto sucede porque en aplicaciones típicas, la salida de un solo modelo oceánico está contenida en la base de datos.

Sin embargo, tenemos un requisito único para tener acceso y almacenar resultados de modelos oceánicos de numerosos proveedores. El tener acceso a múltiples resultados modelos permite la utilización de expertos en modelado de los centros, que se especializan en el desarrollo de modelos. Para acomodar el almacenaje de múltiples modelos de previsiones numéricas, una extensión del concepto de datos de red de Arc Marine fue desarrollada.

La flexibilidad de modelos de datos resultantes del marco de Arc Marine puede ser juzgada en parte por la diversidad de aplicaciones resultantes. Esta diversidad indica que la reutilización de la estructura es posible, con ajustes mínimos a su forma esencial. Por ejemplo, Wright et al. (2007) describen 13 casos de empleo de Arc Marine para usos que cubren asuntos tales como series de tiempo oceanográficas, análisis de litoral, y mediciones oceanográficas químico/físicas. Otros casos de uso examinaron la utilización

de Arc Marine para trazar un mapa del lecho marino (Andrews 2006), la calidad ambiental costera (Brenner y Jiménez 2007), y el rastreo de ballena (Lord-Castillo et al. 2007, 2009).

6.3.5 ISO 19115

Tal y como se ha visto anteriormente en el apartado 3.6, el estándar ISO 19115 para la Información de Metadatos Geográficos consiste en un juego estructurado de objetos de la información que cubre el análisis, el almacenamiento, la distribución y responsabilidades asociadas con la información geoespacial. La ISO 19115 fue diseñada para ser extensible, permitiendo así que aplicaciones específicas de metadatos sean añadidas al perfil base. Por consiguiente, las extensiones de la comunidad-de-interés (COI) han sido creadas para aplicaciones específicas de datos como imágenes y datos en cuadrícula (Xu et al. 2008), más extensos desarrollos de infraestructura de datos espaciales (INSPIRE 2010), examen medioambiental basado en tierra (Peccol 2004), meteorología (WMO 2009) y oceanografía (Australian Ocean Data Centre, 2008).

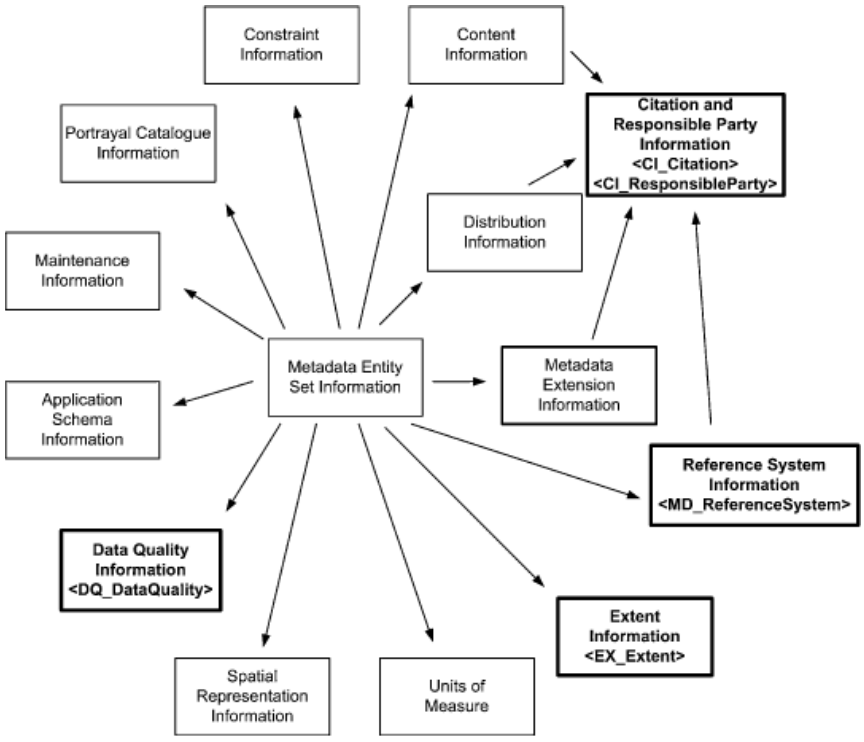

Figura 30: Paquetes de Información contenidos dentro del Estándar ISO 19115 (ISO 2003).

La ISO 19115 completa consiste en 13 paquetes de la información que constituyen el juego de información de metadatos.

La implementación parcial de estos 13 paquetes cumple las exigencias de los investigadores brasileños (p.ej. la transportabilidad de sistema y el historial de procesado de datos). Del paquete de Información de Calidad de Datos, se implementa el aspecto de linaje, permitiendo registrar el historial de procesamiento.

El paquete de información de alcance (magnitud) es utilizado para proporcionar información sobre el área cubierta por el juego de datos.

Las partes responsables del procesamiento y de la información de citado son obtenidos del Paquete de Citación y Parte Responsable. El paquete de Información de Sistema de Referencia proporciona información referente al sistema de referencias usado por el juego de datos. En conjunto, esta implementación se mencionará como Linaje, Extensión, Citación y Referencia (LECR).

6.3.6 Sistema de gestión de base de datos PostgrSQL

PostgreSQL (PostgreSQL 2011) es un sistema de gestión de bases de datos (DBMS) que ha sido aplicado a un amplio surtido de aplicaciones relacionadas con el océano. Beare et al. (2006) y Kupca (2004) utilizaron PostgreSQL en aplicaciones de evaluación de ecosistema, combinando juegos de datos diversos como datos de superficie meteorológica, oceanografía, especies y acústicos en una sola estructura de base de datos. PostgreSQL también ha sido usado como la espina dorsal para un sistema de observación del océano (Fletcher et al. 2008), y en aplicaciones que atraviesan los campos de la medicina (Aloisio et al. 2004), física espacial (Syrjasuo y Donovan 2005), modelado de calidad del aire (Houyoux et al. 2006), oceanografía (Aoyama y 2004 Hirose) y meteorología (Kokkonen et al. 2003).

PostgreSQL sirve como la base para PostGIS (PostGIS 2009), un accesorio que permite la base de datos PostgreSQL espacialmente. Esto quiere decir los tipos de geometría específicos (p.ej. el punto, la línea, el área) relevantes a un GIS, pueden ser incluidos como tipos de datos en la base de datos. Por

consiguiente, una base de datos PostgreSQL con PostGIS puede ser usada como base de datos para apoyar una aplicación GIS. A su vez, el GIS proporciona una amplia gama de posibilidades de visualización de datos, permitiendo incorporar datos con diagramas electrónicos (Goralski y Gold 2007), wikis y portales (Heavner et al. 2011) o esferas virtuales (Ballagh et al. 2011).

PostGIS sigue los estándares OpenGIS® por ser desarrollado por el Consorcio Abierto Geoespacial (OGC 1994). Por seguir los estándares abiertos, la construcción de una base de datos PostGIS permite la adición de otros desarrollos de código abierto que siguen los mismos estándares. Aplicaciones de escritorio de código abierto GIS, tales como uDig (User-friendly Desktop Internet GIS) (Refractions Research, 2009) o Quántum GIS (QGIS) (Quantum GIS Develoment Team, 2011) también adhieren a OpenGIS®. Estos instrumentos proporcionan un surtido de capacidades como la capacidad de visualizar Web Map Services (WMS) y Web Feature Services (WFS) un aspecto importante para la adquisición de datos tanto para el análisis en tierra como en mar (Isenor y Stuart 2007).

6.3.7 El marco de Arc Marine

El modelado de datos es en gran parte considerado una actividad de diseño (Olaya, 2016) donde el pensamiento creativo durante la actividad puede resultar en una diversidad de diseños entre los modelos de datos que son requeridos para objetivos similares.

La aplicación de un marco común al diseño ayuda a reducir esta diversidad. El marco de Arc Marine fue seleccionado como el corazón del diseño, porque este representa uno de los pocos marcos documentados específicos para el contexto oceanográfico.

La diversidad de modelos de datos para aplicaciones similares podría ser abordada si una evaluación objetiva de la calidad de modelos de datos fuera posible. Han habido investigaciones sobre la calidad de modelos de datos (Moody y Shanks 1994, 2003, Moody et al. 2003, Leung y Bolloju 2005, Simsion 2007) que indican desde la perspectiva de usuario, que las calidades

subyacentes importantes son: flexibilidad; integridad; entereza; y comprensibilidad.

Usado como un marco, la integridad y la entereza (Moody y Shanks 2003) de Arc Marine, son ambos difíciles de juzgar. Esto es porque, como un marco, las características dinámicas (p. ej. la integridad) y declaraciones de dominio amplio en cuanto a relaciones de datos (p. ej. la entereza) son dejadas a la implementación específica. La flexibilidad del marco está satisfecha basándose en la diversidad de aplicaciones (previamente apuntadas), mientras que la comprensibilidad es conseguida por el empleo de terminología común. Esto incluye términos tanto oceanográfico como relacionados con GIS.

Por consiguiente, Arc Marine es considerado un marco viable para definir el contenido básico de las entidades de modelo de datos pretendidas para el escenario brasileño, y en última instancia las tablas de bases de datos. El modelo de datos completo desarrollado para esta aplicación consiste en 65 entidades. Más que describir el modelo de datos entero, aquí presentamos las entidades de datos puntuales basadas en el marco de Arc Marine y extensiones para Arc Marine para la inclusión de datos de modelo numérico.

6.3.8 La estructura del Arc Marine

En un contexto oceanográfico, los datos de un perfil vertical consisten en valores de datos discretos en puntos verticales específicos en la columna de agua, típicamente medida como presión (o profundidad) desde la superficie. Como ejemplo, un dispositivo bajado desde el lado de un barco para medir la temperatura del agua, obtendría un perfil vertical de datos. Ejemplos de tales dispositivos serían los termógrafos-de-profundidad fungibles (XBT) (Rual, 1989), perfiladores de la profundidad de temperatura de conductividad con base en barco (CTD) (Cochrane 2007), o perfiladores fungibles CTD (Shi et al. 2011). Los datos recolectados de estos dispositivos pueden ser almacenados como una secuencia de puntos relacionados. Para este tipo de datos, puede ser usada la estructura de punto(puntual) instantánea de Arc Marine.

Las clases relacionadas con la estructura de punto instantánea son mostradas en el Figura 13. El juego de datos es primero descrito usando la clase Paquete de Datos. Esta clase almacena un identificador numérico para el juego de datos. Este identificador es unido directamente a un identificador de mar de prueba, como está contenido en identificador (ID) de navegación en la categoría de navegación. El ID del paquete de datos presente en Paquete de Datos (DataPackage) es renombrado a ID de navegación en la categoría de navegación (en otras palabras, la relación de base de datos une ID de paquete de datos y ID de navegación). La clase de navegación está completamente especificada por Arc Marine.

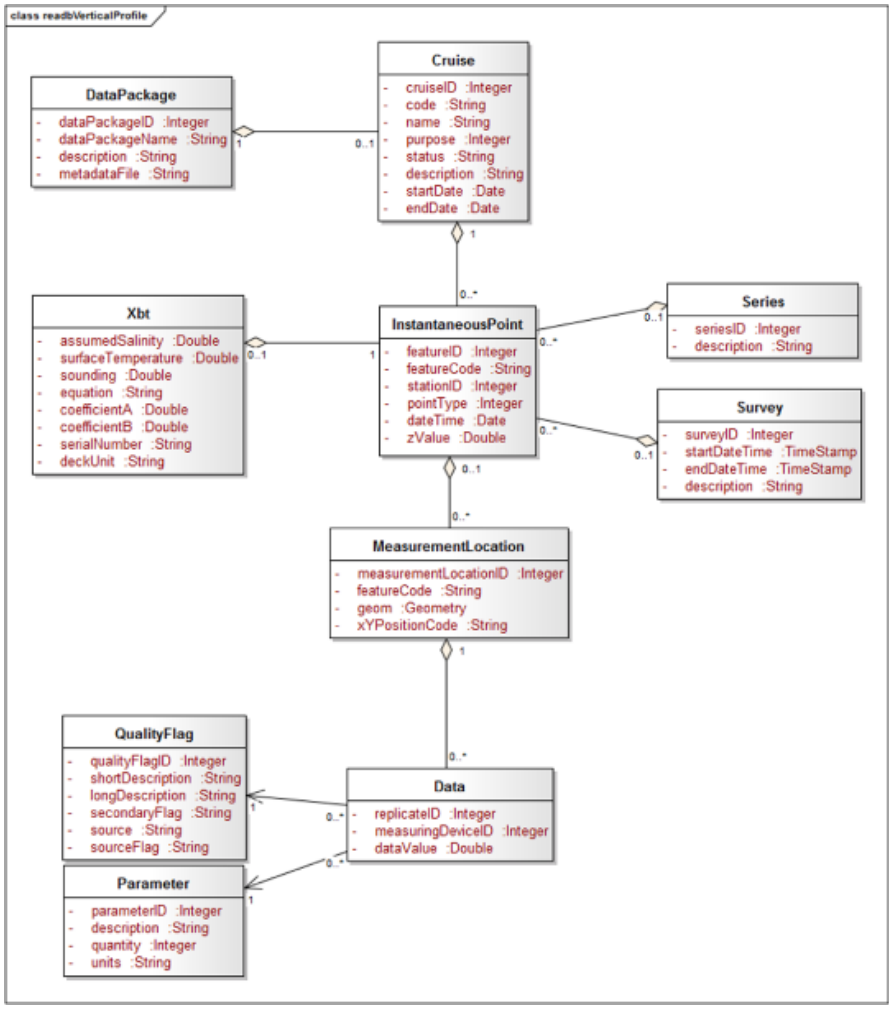

Figura 31: Las clases iniciales utilizadas para datos de perfil verticales

Fuente: Wright (2007)

Hay numerosas relaciones en el Figura que se extienden a clases no mostradas en la figura. Por ejemplo, el registro de navegación específico puede tener trayectos (rutas) asociadas. Un registro de trayecto, contenido en una clase de trayecto (no mostrado), es simplemente una descripción de un tramo del trayecto de un barco, siendo el tramo una línea simple o multi-segmentada.

Múltiples trayectos pueden constituir la travesía (navegación) entera. A lo largo de cualquier trayecto, un vehículo puede ser usado para recoger datos; como contenido en la clase de Vehículo. Un vehículo puede llevar muchos dispositivos (p. ej. instrumentos o sensores) para medir particulares parámetros oceanográficos. Las clases que indican trayectos, vehículos y dispositivos han sido omitidas para mayor claridad.

La clase Punto Instantáneo en el Figura 13 explica el hecho que durante la navegación, la instrumentación realiza mediciones puntuales. Sin embargo, algunos metadatos sólo serán asociados con técnicas perfiladoras específicas. Para datos XBT, estos metadatos están en la clase Xbt. Asimismo otro dispositivo de clases específicas de metadatos (no mostrado) existirían para metadatos relacionados con modelos CTD, mediciones de pérdida de transmisión, etc.

A cada punto en InstantaneousPoint le es asignado un identificador de características. Este identificador es utilizado en la relación a la clase Localización de Medición, donde son almacenados los valores de x, y, z para todas las medidas. Los valores de x,y,z son codificados dentro de una columna PostGIS geom (geometría), para facilitar su empleo por aplicaciones de clientes informados de PostGIS. El campo de geometría contiene una codificación del x, y (longitud, latitud) posición, y opcionalmente el componente posicional z. A cada juego de medidas en un solo x,y,z es asignado un ID (identificador) de locación de medición. La relación de la base de datos a la clase de Datos establece una conexión entre identificador de locación de medición en Medición de Locación y el campo de identificador de locación de medición en la clase de Datos.

Los valores de datos reales serán almacenados en el atributo de Valores de datos dentro de la clase de Datos. El tipo de datos es identificado usando ID de parámetros. Cada parámetro ambiental (p.ej. temperatura del agua, salinidad y contenido de oxígeno del agua), es descrito usando los registros en la clase de Parámetro. El campo de duplicar ID es simplemente un contador para identificar mediciones duplicadas del mismo parámetro. Las marcadores de calidad pueden ser asignadas a valores individuales utilizando los marcadores de calidad almacenados en la clase Marcadores de Calidad. Los marcadores de calidad son una extensión para Arc Marine.

Para la comunidad oceanográfica, las características útiles de este modelo provienen del acoplamiento débil de las entidades. Tal acoplamiento, reconocido por el empleo de números de identificador (p.ej. atribuye nombres que incluyen ID) permite diversidad en el embalaje de los juegos de datos en unidades (p.ej. Clase de paquete de datos en lo alto del gráfico), por la calidad y definiciones de parámetros (p.ej. ver la clase de Parámetro al final del gráfico).

También, los metadatos específicos al tipo de perfil son mantenidos vía una clase específica (p.ej. clase Xbt en el gráfico).

Esto permite diversidad en las mediciones al tiempo que permite una fácil expansión a perfiles recogidos desde otros dispositivos.

6.3.9 Extensión específica de aplicación - Datos Modelos Numéricos

Un uso de la base de datos de evaluación rápida ambiental será el almacenaje de datos requeridos para pronosticar condiciones acústicas debajo del agua. Para obtener las estimaciones de condiciones acústicas en futuras ocasiones, se utilizan datos de salida medioambientales de modelos numéricos de pronóstico oceánico. Tales modelos numéricos producen las previsiones regulares de temperatura del océano y salinidad a través de toda la columna de agua, que puede entonces ser usada para pronosticar las condiciones acústicas.

El marco de Arc Marine contiene estructuras para el almacenaje de modelos de datos numéricos. Sin embargo, esta aplicación es suficientemente

diferente(variada) para garantizar la construcción de clases especializadas. Estas clases amplían la funcionalidad del marco de Arc Marine para satisfacer las necesidades específicas del equipo de investigación que realiza el modelado acústico. Expresamente, los modeladores acústicos desean comparar predicciones acústicas cuando utilizan entradas (de datos) de modelos oceanográficos ambientales independientes, con las condiciones acústicas observadas. Las clases resultantes se muestran en el Figura 14.

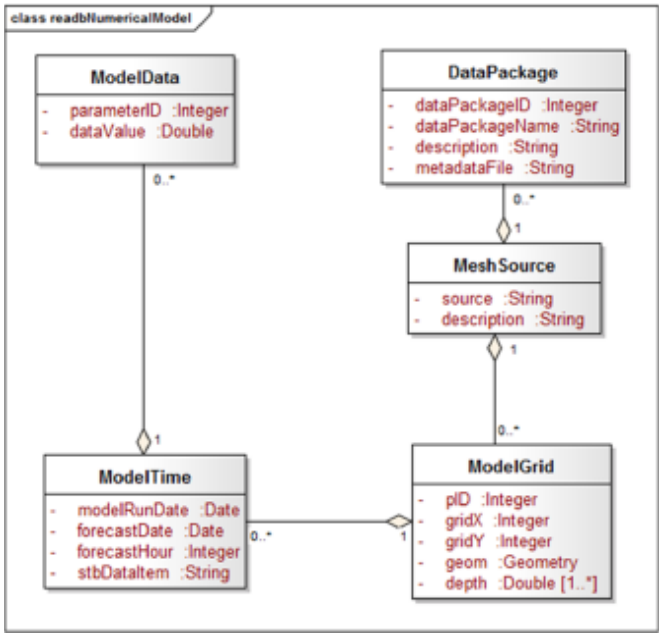

Figura 32: Los datos de modelos numéricos oceánicos son almacenados en clases que son extensiones para Arc Marine. Las clases numéricas de modelos oceánicos permiten el almacenaje de resultados de simulación provenientes de múltiples modelos numéricos del océano, todas proveen múltiples previsiones para cada ejecución de modelos numéricos o tiempo de ejecución.
Fuente: Wright, et. al. (2007)

Hay tres clases específicas para los modelos de previsiones numéricas oceánicas, con dos clases de soporte. Las clases de soporte son MeshSource (fuente de red) (Figura 14) y ModelAsset (modelo activo)(no mostradas). MeshSource contiene conjuntos de datos de metadatos representados por un identificador de nombre simple para un modelo numérico de región específico. Esta clase es efectivamente un inventario de juego de datos, con el nombre identificador que es usado en el modelo de datos, indicando las salidas de un modelo numérico específico para una región geográfica específica. Esta clase también puede ser usada para nombrar cualquier red asociada con datos regulares (promedio) (p. ej. tablas no mostradas), como promedios

climatológicos. La clase ModelAsset proporciona el enlace entre el llamado modelo numérico y el identificador de paquete de datos, utilizado para rastrear el completo juego de datos asociado con el modelo numérico.

Las principales clases para las previsiones oceánicas de modelo numérico, están contenidas en ModelGrid (modelo de red), ModelTime (modelo de tiempo) y ModelData (modelo de datos). ModelGrid contiene la especificación de la red para la fuente particular de modelo numérico. Los atributos de gridX y gridY indican los índices de x e y de la casilla (celda) modelo. El componente vertical es almacenado en el campo de profundidad. Esta implementación específica utiliza el tipo de almacenaje de serie de PostgreSQL DBMS.

Esto permite el almacenaje de una serie dentro de una sola fila de tabla de base de datos, en un campo llamado profundidad.

La clase ModelTime es utilizada para contener los detalles de la ejecución del modelo numérico y el tiempo de previsión. Los modeladores acústicos desean acceder, utilizar, y evaluar múltiples modelos numéricos oceanográficos, lo que impone el requisito de muchas marcas de tiempo para distinguir las previsiones. La fecha de ejecución del modelo numérico es usada para distinguir entre las muchas ejecuciones de modelos numéricos. La fecha de pronostico y el tiempo son entonces usados para identificar específicamente los datos pronosticados. Finalmente, el atributo stdDataItem proporciona un enlace a la aplicación de medioambiente Cama de Prueba de Sistema (STB; Johnson 2007), usado para el modelado acústico.

Los detalles de esta estructura de clase se comprenden mejor utilizando un ejemplo. Esta estructura fue utilizada para almacenar modelos de datos numéricos adquiridos de la Canada-Newfoundland Operational Ocean Forecasting System (C-NOOFS) (Dombrowsky et al. 2009, Industria pescadera y Océanos Canadá 2008). C-NOOFS tenía datos numéricos disponibles de dos dominios de modelos numéricos diferentes; un dominio con una resolución de cuadrícula de 1/4° y un segundo con resolución de 1/12°. Estos dos modelos numéricos fueron designados como fuentes CN04 y CN12. En cualquier día en particular, los resultados numéricos generados por

C-NOOFS fueron obtenidos a bordo del barco vía una conexión de satélite. Cada modelo numérico produjo una previsión que incluyó la temperatura de agua y campos de salinidad en intervalos de seis hora durante los siguientes cinco días. Nótese que para una descarga específica, la ejecución de hoy de una previsión del futuro cuarto día, representa el mismo tiempo de previsión que la ejecución de mañana de una previsión para el futuro tercer día. Así, todos requieren el nombre del modelo numérico de la fuente, el tiempo de ejecución del modelo numérico, la fecha pronosticada y la hora para identificar únicamente la previsión específica.

La clase final es ModelData. Esta clase contiene los datos del modelo numérico del océano. Una vez más, el nombre de la fuente, el tiempo de ejecución del modelo numérico, la fecha de previsión, y el tiempo de previsión representan la clave primaria para la clase. El parámetro que es almacenado, es identificado utilizando el atributo de identificador de parámetro de Arc Marine, mientras que el valor real está contenido en Valor de datos.

Para los investigadores acústicos, este conjunto de clases permite una considerable diversidad en los modelos de datos oceánicos gestionados. Múltiples cuadrículas de múltiples modelos oceánicos, cada una suministrando pronósticos que cubren diferentes períodos en el futuro, pueden ser recopilados en esta estructura de clase.

6.3.10 Linaje, Extensión, Citación y Referencia (LECR) de la ISO 19115

El linaje es el término que describe la procedencia o la historia asociada con un juego de datos. La información de linaje es una forma de metadatos que describe la fuente y los procesos administrativos utilizados para producir o modificar el juego de datos. El linaje de metadatos guarda un recuento de la historia del procesamiento desde la recolección de datos al producto final.

Esta información de procesamiento histórico es un importante componente de la calidad del conjunto de datos. El historial proporciona a los usuarios de conjuntos de datos, de la capacidad de evaluar la totalidad del tratamiento (procesamiento), e identificar temas con los métodos de procesamiento,

coeficientes usados en el procesamiento, o aquellas personas implicadas en el procesamiento.

Este tipo de metadatos proporciona a los usuarios de la información requerida para comprender mejor el juego de datos, incluyendo la fiabilidad y la utilidad del juego de datos. Estas características son factores en la generación de confianza (Adams et al. 2003) que es un aspecto importante de la siguiente generación del World Wide Web (Mahmood 2007).

El modelo de linaje presentado aquí representa un subconjunto de la ISO (2003) 19115 para el linaje de metadatos y se muestra en el Figura 16. El linaje comienza con un identificador dentro de la clase LI_LINEAGE. El identificador mantiene la relación para una línea de historia específica (p. ej. el linaje). Este linaje puede ser aplicado a uno o muchos juegos de datos, como esta indicado por la relación a LI_SOURCE. LI_Source, proporciona la información referente a juegos de datos individuales en el paquete de datos. Esta clase es el equivalente a la ISO 19115 elemento *92. El identificador único liSourceID representa una sola fuente de datos. La secuencia de identificadores liLineageID en la clase LI_SOURCE representa entonces una serie de alteraciones a los datos. Cada fuente, como esta representado solo por liSourceID, tiene una descripción, denominador de escala e identificadores para el sistema de referencia, citas y alcance geoespacial y temporal.

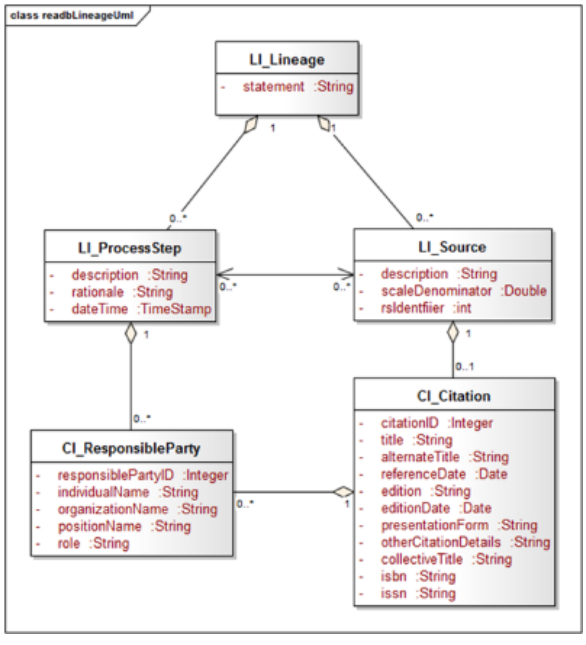

Figura 33: El estándar ISO (2003) para metadatos geográficos es usado para definir un conjunto de clases utilizadas para el proceso de rastreo de un juego de datos. Entidades unidas, presentes en el modelo relacional, proporcionan el enlace entre los elementos de ISO 19115

Fuente: Wright (2007)

Un linaje puede ser construido desde una secuencia de pasos del proceso. Dentro de la base de datos REA, los informes de linaje son almacenadas como un atributo de linaje. Los pasos de proceso de linaje son almacenados dentro de una tabla de detalle enlazada al linaje. Esto permite a cada paso de proceso ser almacenado separadamente, pero también permite que cada acción principal dentro de un paso del proceso, sea marcada en el tiempo individualmente. En el modelo relacional, esta relación es mantenida en la entidad JO_Process_Step_Lineage (nótese que JO indica una entidad de unión (join) y no se muestra aquí).

Los pasos de proceso que son aplicados a los datos son descritos en la clase LI_ProcessStep. Los pasos del proceso son numerados utilizando el identificador único de liProcessStepID. Cada paso de proceso tiene una descripción, un fundamento y una fecha de aplicación.

Una persona con algún papel en los pasos del procesamiento individual puede ser identificada en la clase CI_ResponsibleParty (grupo responsable).

Los ejemplos de roles incluirían a la persona responsable de la creación del paso de procesamiento, o la aplicación del paso de procesamiento.

La clase CI_CITATION permite al almacenaje del metadatos asociados con un informe, artículo o publicación. El campo de citationID únicamente identifica la cita y también enlaza la cita a el liSourceID. Esto quiere decir que la cita puede ser vinculado al juego de datos usado en la publicación. La inclusión de citas a publicaciones dentro de la base de datos permite al rastreo tanto de fuentes de datos como de productos de datos. Los metadatos de citación incluyen el título, el titulo alternativo, la fecha de referencia, la información de edición, la forma del material de presentación, detalles de cita y números estándar internacionales para libros o series. Este juego de metadatos imita en conjunto la ISO 19115 elemento *359.

6.3.11 Salidas de usuarios

En algunos casos es útil incorporar y utilizar un activo de datos creado y gestionado por una fuente externa. Un ejemplo de tal activo de datos sería los datos de batimetría de resolución de dos minutos, proporcionados por el Banco Nacional de Datos Oceanográficos (BNDO) brasileño. Típicamente, tal activo de datos es manejado por la fuente externa a través de una serie de actualizaciones, con actualizaciones proporcionadas en una forma constante o estándar.

Aunque fuera posible incorporar el activo de datos directamente en la base de datos, manejar el recurso externo a la base de datos proporciona ciertas ventajas. Primero, el tamaño de la base de datos no se aumenta debido a la incorporación del activo externo. Segundo, ya que la forma original del activo externo es mantenida dentro del sistema, la actualización abre camino a la suma de activos para un reemplazo de archivo (asumiendo que la actualización mantiene el formato del activo original).

El concepto de salidas de usuario es introducido como el medio para manejar tal activo de datos externos. Las salidas de usuarios permiten al software de usuario ser conectado al DBMS, similar a las funciones que existen dentro del DBMS. Esta capacidad es en particular útil para tener acceso a juegos de datos externos en una manera compatible con el estándar SQL (Digital Equipment Corporation 1992).

Una salida de usuario implica la creación de un enlace entre un activo de datos externo y el sistema de gestión de datos. La gestión de este enlace es controlada por el DBMS. Efectivamente, el DBMS es utilizado para dirigir el software de procesamiento al activo de datos externo y para controlar que software de procesamiento es usado para conseguir el activo de datos.

Para el activo ETOPO2, una salida de usuario fue creada usando una combinación de PL/PerlU y Java. El código PL/PerlU, que es llamado loadEtopo2Table, existe dentro del DBMS. Este código realiza comprobaciones de errores simples en los parámetros suministrados por el usuario, representando estos parámetros los límites de latitud/longitud

superiores izquierdos e inferiores derechos de un cuadro geoespacial. Este código también maneja la interfaz al software Java.

El software Java acepta las coordenadas del cuadro delimitador e interactúa directamente con el formato de archivo del activo de datos externo para obtener los datos. El software extrae una región de datos de batimetría del activo ETOPO2 y coloca los datos extraídos en una tabla dentro de la base de datos. La salida de usuario es iniciada vía un comando tipo SQL, como se muestra debajo:

SELECT loadEtopo2Table (45.0,-64.0, 40.0,-60.0);

El uso de la sentencia anterior produce la creación de una tabla batimétrica que contiene los valores de profundidades ETOPO2 en la región descrita. El acceso a estas profundidades utilizando QGIS, incluyendo una línea de costa, y el color de codificación de los puntos de acuerdo con la profundidad, derivan en la Figura 22:

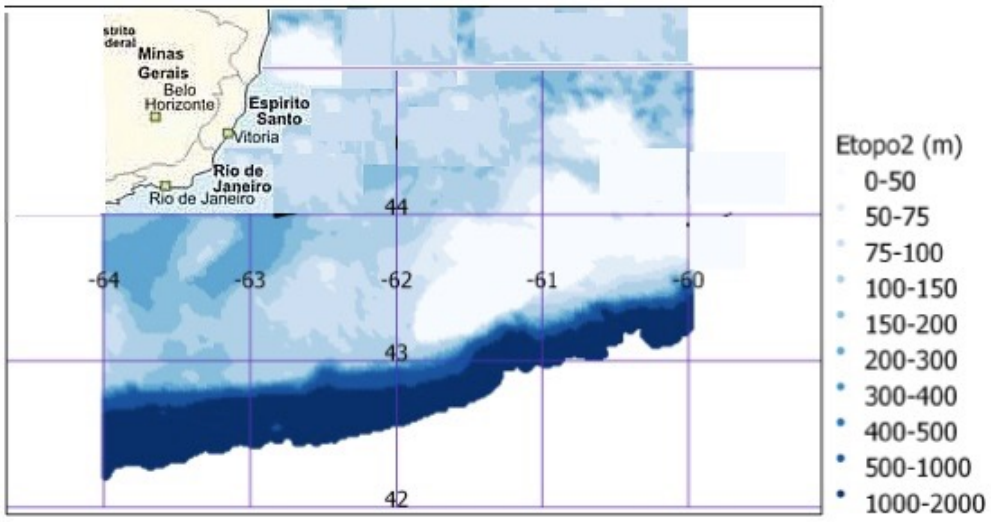

Figura 34: Puntos de profundidad obtenidos de la función de salida de usuario. Las líneas de cuadrícula indican la latitud en grados al norte y la longitud en grados al este

6.3.12 Análisis general del modelo

El actual sistema de gestión de datos oceanográficos en Brasil es usado extensivamente en un entorno de laboratorio, es decir, sin posibilidad de captura instantánea o conexión de los datos comunes en repositorios integrados. El sistema de gestión de los datos, como el que hemos descrito aquí representa la Versión 2 (V2) del sistema Arc Marine. La primera versión

(V1), tenia como énfasis la aplicación tanto en tierra como en mar, mientras la V2 hasta ahora sólo ha sido usado en el laboratorio (Wright, 2007). El diseño de la Versión 2 incorpora Arc Marine e ISO 19115, al tiempo que incorpora las clases de modelo numérico V1. El concepto de salida de usuario era también una adición del V2.

La introducción de Arc Marine en V2 era en respuesta a un requisito puesto en V1 para mejorar la flexibilidad al incorporar nuevas fuentes de datos de punto. La estructura común de Arc Marine para datos de punto (Figura 22) cumple con este requisito. La clase InstantaneousPoint permite la incorporación de datos de punto diversos en la base de datos a través de la adición de la definición de tipo de parámetro apropiada para la medición de punto específica. Por ejemplo, incorporar datos de sedimento del océano obtenidos de instrumentación desarrollada (Osler et al. 2006) requiere de la adición del tipo de parámetro y una clase de metadatos para apoyar los metadatos del instrumento específico. Sin embargo, la limitación de este enfoque es la definición de la clase de metadatos para apoyar la instrumentación específica.

El empleo histórico de Arc Marine ha tenido lugar en el entorno Esri GIS. Sin embargo, hemos mostrado que el marco de Arc Marine puede ser usado para construir un sistema de información a un entorno non-Esri y perfectamente aplicable como modelo de gestión de datos oceanográficos en Brasil. También, hemos mostrado que Arc Marine es suficientemente flexible para permitir a la separación de juegos de datos internos y externos como fue demostrado por el método de inclusión del juego de datos batimétrico. El concepto de salida de usuario proporciona el mecanismo para esta separación. Esto es similar en concepto a un Servicio de Rasgo de Web que tendría acceso a un recurso de datos remoto, manejado por la autoridad remota.

Sin embargo, la actual estructura de datos realmente sufre de poca utilización de tipos de geometría complejos. La estructura actual utiliza el punto geoespacial de 2 dimensiones y estructuras de red donde la tercera dimensión (p. ej. z) es independiente del tipo de datos geoespacial. La utilización de la geometría de punto 3-dimensional proporcionaría productividad cuando esté conectada a instrumentos GIS capaces de utilizar

estos tipos de geometría y mejoraría la capacidades de procesamiento y visualización tales como la representación de la superficie.

La capa de información también se beneficiaría de una investigación en el empleo de la clase de Arc Marine que trata con animaciones y fotografías. Serían aquí de interés particular las referencias geoespaciales en transmisiones de vídeo. Esto sería de interés en ambos entornos marino y terrestre.

El acercamiento de enfoques estándar de base utilizado aquí para la capa de información del sistema de gestión de los datos es esencial si la interoperabilidad de juegos de datos entre sistemas debe ser alcanzada. Sin embargo, se debe notar que la interoperabilidad de estructuras de datos no necesariamente aborda la interoperabilidad semántica, lo que es requerido para consolidar activos de datos a través de múltiples laboratorios (Graybeal et al. 2012).

El estándar ISO 19115 contribuye en aquellos aspectos relacionados con el linaje o el historial de procesamiento de un juego de datos. Aunque las aplicaciones de base de datos de ISO 19115 hayan sido construidos por otros, la armonización de componentes de ISO relevantes en un modelo de datos emparentados basado en el marco de Arc Marine es única. También, la utilización de salidas de usuarios para incorporar juegos de datos externos en el sistema de gestión de datos representa un método para utilizar eficientemente grandes juegos de datos que son manejados externamente.

Este trabajo ha mostrado que las estructuras de datos dentro del marco de Arc Marine proporcionan la flexibilidad para el procesamiento de datos oceanográficos en la costa y en el mar, fuera del entorno Esri. Arc Marine es también lo suficientemente flexible como para permitir la combinación del marco con otros estándares, como la ISO 19115.

Combinado con el concepto de salida de usuario, también hemos mostrado que Arc Marine permite la separación de juegos de datos internos y externos. Esta separación significa que la gestión de los juegos de datos sigue siendo responsabilidad del autor y no necesariamente responsabilidad del sistema de gestión de datos oceanográficos.

La utilización de datos históricos y metadatos asociados es importante para la comunidad oceanográfica, pero es también importante desde la perspectiva de la rentabilidad del laboratorio, dado que la recolección de datos en el mar es costosa. Sin embargo, la manera más eficaz de organizar los datos para permitir un acceso y uso eficiente, sigue siendo un problema para la comunidad oceanográfica (Bechini y Vetrano 2013, McCann y Gomes 2008). La capacidad de manejar una colección de datos diversos dentro de una sola infraestructura de información es una tarea compleja para un país con las dimensiones que tiene Brasil

La propuesta que presentamos se trata de un modelo para manejar correctamente los datos oceanográficos de los investigadores brasileños, de manera que fomenten su utilización dentro de la comunidad de usuarios. Con esto se combina la premisa de que la implementación de especificaciones abiertas es importante para identificar los datos y metadatos necesarios para ser gestionados. El trabajo reciente por Aliprandi (2011) sugiere que el empleo de estándares abiertos, es también un paso positivo hacia la interoperabilidad.

La interoperabilidad es reconocida como un componente crítico cuando sostiene o asiste las necesidades de la comunidad más amplia (Baker and Chandler 2008).

En el nivel de laboratorio, este trabajo también aspira a mejorar la consistencia de datos y reducir la duplicación de esfuerzos de los usuarios. Cuestiones referentes a la consistencia de datos aparecen cuando los análisis del usuario producen múltiples versiones de juegos de datos.

La duplicación de esfuerzos del usuario es una cuestión cuando cada usuario crea sus propias aplicaciones para ejecutar tareas que son comunes a muchos usuarios; por ejemplo, para leer y analizar el formato de salida de datos de un instrumento.

El deseo de gestionar correctamente la colección de datos provoca varias preguntas relacionadas con la estructura dentro de la cual los datos son manejados. Estas preguntas se relacionan con cómo los datos deberían ser organizados para promover estructuras de datos comunes, interoperabilidad, y aplicaciones de tratamientos comunes, al tiempo que permitan flexibilidad

para gestionar los nuevos tipos de datos que surgen cuando se desarrolla instrumentación en-mar. Las estructuras de datos también deberían promover la convergencia de las estructuras dentro de la comunidad que las emplea, a través de la utilización de estándares y especificaciones existentes, y también por la identificación de los componentes principales de la información requeridos por un sistema geoespacial para apoyar la gestión de datos.

6.4 Políticas científicas

La planificación de la gestión de datos es una parte muy importante de la investigación. Varios organismos internacionales de financiación requieren que los investigadores planifiquen adecuadamente a sus datos. Es particularmente importante para facilitar el intercambio de datos, asegurando la sostenibilidad y la accesibilidad de los datos a largo plazo, y permitiendo que los datos pueden volver a utilizar para la investigación futura.

Para la gestión eficaz de los datos oceanográficos, la planificación debe comenzar cuando la investigación está siendo diseñada y debe tener en cuenta tanto cómo se gestionarán los datos durante la investigación y la forma en que se compartirán después. La gestión de datos no es sólo responsabilidad del investigador que ha creado o recogido los datos. Varias partes están involucradas en el proceso de investigación y pueden desempeñar un papel en asegurar datos de buena calidad, la protección de ellos y facilitar el intercambio de datos. Es crucial que las funciones y responsabilidades se asignan y no sólo presume. Para la investigación en colaboración, la asignación de funciones y responsabilidades entre los equipos de investigadores es importante. Esto implica pensar críticamente en cómo los datos de investigación pueden ser compartidos, lo que podría limitar o prohibir el intercambio de datos, y si es posible tomar medidas para eliminar esas limitaciones.

Para que los investigadores depositen los datos con una eficiente descripción sistemática, ellos deben tener en cuenta los siguientes aspectos para gestionarlos y difundirlos adecuadamente:

6.4.1 Plan de gestión

Es fundamental que los investigadores desarrollen un plan de gestión de datos para sus proyectos, para ser presentada ante las instituciones que poseen vinculo y órgano de financiación. Un plan de gestión de dato notifica a los organismos de financiación (por ejemplo, las fundaciones brasileñas de amparo a investigaciones) acerca de las intenciones para los datos de investigación. Dado que estos detalles pueden cambiar durante el curso de su proyecto, un plan de gestión de datos debe ser considerado como un "documento vivo" de los cuales es probable que sean por lo menos dos versiones: una en el inicio del proyecto y una actualizada al finalizar el proyecto.

Los detalles generales de un plan de gestión de datos pueden ser verificados en la secuencia (más detalles puede ser revisados en el apartado 2.5 - Plan de gestión de datos, p.75) .

1 ¿De que se trata? (Lo que es el proyecto de investigación?)

2 ¿Qué es? (Formatos, tamaño aproximado, método recopilación o producción, etc.)

3 ¿Quién es responsable de la misma y quién es el propietario?

4 ¿Dónde estará depositado (incluyendo copias de seguridad)?

5 Si va a ser accesible para su reutilización, y si es así cómo, cuándo y por quién?

6 ¿Cuánto tiempo debe mantenerse?

La forma más eficiente de completar un plan de gestión de dato es mediante el uso de una herramienta de planificación de la gestión de datos de la institución en que la investigación tiene vínculo. En el caso que no haya una herramienta especifica, es recomendable el uso del Data Management Planning Tool (Digital Curation Centre, 2015). Una vez que haya completado un plan de gestión de datos, puede adjuntar registros de datos descriptivos. Los registros de datos en las pertenencias son los registros de catálogo. Es decir, se dan algunos detalles mínimos sobre los datos para registrar su ubicación (que también pueden vincular directamente a los datos). Los registros de datos se

pueden mantener en privado, compartidos solamente con los investigadores especificados, o puede elegir para publicar en el Banco Nacional de Dados Oceanográficos (BNDO) de Brasil.

6.4.2 Disponibilidad de los datos

Datos de investigación financiados con fondos públicos son un bien público, producido en el interés público, que deberá hacerse abierto y accesible con el menor número posible de restricciones en forma oportuna y de manera responsable que cumple con un estándar ético y no viola la propiedad intelectual.

Aunque no sea obligatorio, se recomienda a los investigadores financiados por agencias de financiación ofrecer copias de los datos creados o reutilizados durante su investigación. Las principales agencias recomiendan el deposito de los *datasets* que sustentan las investigaciones, entre ellas la Fundação de Amparo à Pesquisa do Estado de São Paulo (FAPESP) y el Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq).

Los investigadores deben asegurarse de que cuando publican resultados de la investigación, los datos que apoyan estos resultados se depositan al mismo tiempo con un depósito digital responsable. Deben hacer referencia a la publicación, donde sea posible, a través de una citación formal, encontrar y acceder los datos que sustentan los resultados presentados, mientras sólo deben publicar datos de la investigación en repositorios que proporcionan identificadores persistentes para los datos (hemos tratado este tema en el apartado 2.6.9 - Los identificadores).

6.4.3 El deposito de los datos

Los datos pueden ser depositados en los repositorios institucionales o repositorios responsable correspondiente. En cualquier caso los metadatos deben ser proporcionadas al BNDO mediante la creación de un registro.

La mayoría de los esfuerzos de recopilación de datos a gran escala ahora implican el uso de softwares integrados y programas especiíficos, pero todavía

hay situaciones en las que se requiere la entrada de datos - por ejemplo, la introducción de los registros administrativos, datos de observación, o respuestas a las preguntas abiertas. Una serie de herramientas de software están disponibles para hacer más fácil la tarea de documentación. Para proyectos que requieren la entrada de datos directamente de las observaciones marinas, una variedad de programas no sólo hará que la entrada de datos sea más fácil, sino también llevará a cabo comprobaciones de integridad de los datos que se introducen creando instrucciones de programación para leer los datos.

El almacenamiento adecuado de los datos de la investigación es esencial para garantizar su seguridad y la seguridad. Existen una gama de soluciones de almacenamiento para los datos de la investigación, pero, en particular, es posible archivar en herramientas de estilo 'DropBox' que permite a un grupo de investigación almacenar y compartir datos de forma rápida y segura. Este servicio es fácil de usar y el acceso al igual que otras ofertas de nube pública, pero está alojada con seguridad en servidores con gran capacidad. En el caso que el proyecto tenga grandes necesidades de almacenamiento o especializados, o mismo la necesidad de confidencialidad, es recomendable el uso de un repositorio institucional.

Todos los datos creados o reutilizados durante una investigación deben estar disponibles para su reutilización lo más breve posible al final de la subvención.

Indicar cómo y dónde se va a almacenar copias de los archivos de investigación para garantizar su seguridad, así como el número de copias que va a tener y cómo va a sincronizarlos. La mejor práctica para la protección de datos es almacenar múltiples copias en varias ubicaciones.

6.4.4 El formato de los datos

Aunque no exista una estructura universal de datos oceanográficos, los formatos presentados en el apartado 3.4 son los más utilizados en investigaciones marinas. Estos formatos integran la estandarización internacional y su uso facilita el intercambio entre repositorios. Además, es necesario definir la estructura y la dimensión que tienen los datos,

describiendo la cantidad y el tamaño de los ficheros, si existen ficheros secundarios y la organización y el nivel de detalle que tienen.

Es muy importante describir los formatos de los datos en la presentación, la distribución y las fases de conservación (teniendo en cuenta que estos formatos pueden ser los mismos). La selección de formatos para archivar pueden hacer el procesamiento y la liberación de información más rápida y más eficiente. El investigador también debe asegurarse que los formatos de datos oceanográficos que no sean comunes en el repositorio de deposito y que no sean propietarios podrán utilizarse en el largo plazo.

Los datos oceanográficos son muy diversos y normalmente son organizados en variados ficheros dentro de un solo conjunto de datos y en otras será necesario organizar los datos en varios conjuntos diferentes. El investigador debe definir el sistema de relaciones entre los distintos componentes del conjunto de datos.

En situaciones que el conjunto de datos tenga una identidad específica muy concreta y su estructura se organiza en una multiplicidad de registros, es recomendable enviar una copia del fichero principal al Banco Nacional de Dados Oceanográficos (BNDO) de Brasil con la referencia para localizar los ficheros secundarios.

Es necesario tener en cuenta que al escribir una propuesta de investigación con la previsión de generar datos, es útil pensar en los "datos" en el sentido más amplio, incluyendo los archivos de datos numéricos, transcripciones de entrevistas y otros materiales cualitativos tales como diarios y notas de campo. Cada vez más, los datos de investigación oceanográfica incluyen formatos de audio y vídeo, datos geoespaciales y muchos repositorios de datos requieren la descripción de todos los ficheros vinculados al archivo principal. El archivo y la difusión de conjuntos de datos derivados, es decir, que resultan de la combinación de más de una fuente de datos, incluyendo los datos existentes fuera del ámbito actual de la investigación, también deben ser considerados (ver el apartado 2.3.5 - Formatos de archivos de los datos, para una discusión más en profundidad).

6.4.5 Metadatos

Los datos de investigación deben ir acompañadas de metadatos para proporcionar a los usuarios información esencial para entender los datos de forma independiente y permitir la reutilización científica. La documentación debe describir por lo menos el origen de datos, trabajo de campo y métodos de recolección de datos, procesamiento y/o el investigador responsable. Elementos de datos individuales como variables o transcripciones deben estar claramente identificado. Al depositar los datos con un repositorio responsable, es necesario crear un registro de metadatos de acuerdo con la normas de este repositorio y siguiendo un estándar de metadatos que explican referencias como el propósito, origen, tiempo, ubicación geográfica, creador(s), las condiciones de acceso y las condiciones de uso de los datos. (A veces se requieren múltiples estándares de metadatos). El repositorio utilizará estés metadatos para publicar, difundir y promover los datos de la investigación. Todas las publicaciones basadas en los datos resultantes de una investigación deben incluir específicamente información sobre dónde y cómo se puede acceder a los datos, idealmente a través de una citación formal.

6.4.6 Identificadores geográficos y datos geoespaciales

Algunos proyectos en el campo de la oceanografía recogen datos que contienen identificadores geográficos directos e indirectos que se pueden codificar de forma geográfica y usar con una aplicación de mapas. Identificadores geográficos directos son direcciones reales (por ejemplo, de un incidente, localización de puerto, uno laboratorio, etc.). Identificadores geográficos indirectos incluyen información de ubicación, tales como una zona de pesca, localización de un satélite, una embarcación, etc, y el lugar en que el demandado se crió.

Se anima a los investigadores a añadir a las variables del conjunto de datos derivados que agregan sus datos a un nivel espacial que puede proporcionar un mayor anonimato sujeto (tales como estado, país, división o región). Es deseable que los productores de datos de dirección geográfica puedan coordinar los datos ya que a menudo pueden producir mejores tasas de

geocodificación con su conocimiento de la zona geográfica. Cuando los productores de datos conviertan las direcciones de las coordenadas geoespaciales, más adelante se pueden agregar los datos a un nivel superior que protege el anonimato demandado. En tales casos, los identificadores geográficos originales deben ser guardados en un archivo de datos separado que también contiene una variable para enlazar a los datos de la investigación. El archivo con los identificadores directos debe ser presentado al archivo en separado. Es recomendable que los investigadores verifiquen los requisitos para la presentación que contiene la información geográfica detallada.

7 CONCLUSIONES

Después del análisis del contexto internacional y de la situación en Brasil en el ámbito de la gestión de datos oceanográficos que se ha llevado a cabo en los capítulos precedentes vamos a sintetizar las principales conclusiones que se derivan de los cuatro objetivos que nos habíamos planteado

Gestión de datos de investigación

Cada vez se pone mayor énfasis en la gestión de datos de investigación en el entorno a su volumen ascendiente y nuevas infraestructuras para su preservación, lo que está impulsando a las instituciones académicas a desarrollar y desplegar nuevas iniciativas. El análisis de las necesidades de datos de los investigadores pone en evidencia nuevas demandas y condiciones para conectar conocimientos y descubrimientos del entorno científico, tanto formales e informales para compartir, analizar, y reutilizar datos. El análisis sobre la gestión de datos de investigación es uno de los retos futuros que la comunidad científica deberá asumir para su avance. Se trata de una nueva manera de organizar la información y que requiere esfuerzos importantes en el aprendizaje de nuevos métodos de trabajo y colaboración con los agentes implicados.

Sin embargo, las nuevas iniciativas de gobiernos y comunidad científica para el desarrollo de nuevas infraestructuras para preservación y acceso a los datos de investigación han avanzado al punto de establecer estándares internacionales que posibilitan la interoperabilidad de datos y cooperación técnica. El aumento de la capacidad tecnológica para procesar gran cantidad de datos ofrece nuevas oportunidades y beneficios fundamentalmente para tres sectores: gobierno, comunidad científica y la sociedad en general. Conocer y alfabetizar sobre estas cuestiones requiere un esfuerzo de todos los investigadores involucrados para fomentar una mentalidad sobre la importancia de estos datos y la cultura de análisis, ya que se trata de la adopción de las nuevas tecnologías.

Tal como en todas las áreas del conocimiento, las ciencias marinas ahora pasan por una transformación en la manera de cómo operan los datos de

investigación. En el caso de Brasil, las fallas estructurales en la gestión de datos oceanográficos e investigaciones realizadas en el entorno de los océanos aumenta la necesidad de mecanismos eficientes de divulgación científica, especialmente con respecto a la disociación de los datos válidos y útiles, de los contenidos innecesarios u obsoletos para investigadores, centros de investigación y universidades. La organización adecuada de los datos en los ecosistemas marinos implica varios pasos, desde la adquisición hasta su archivo, control de calidad y su posterior difusión.

Entre las cuestiones que se ha planteado en esta tesis está la diferencia entre ¿Cuál es la actual situación brasileña en la gestión de los datos de investigación oceanográfica?, ¿Qué alternativa puede aportar una solución para optimizar la actual infraestructura de modo que atienda los estándares internacionales? Este trabajo ha tratado de abordar estas cuestiones.

Situación Internacional

En relación al escenario internacional, el panorama europeo es demasiado complejo y fragmentado y esto es un obstáculo para lograr un impacto positivo como modelo para las crecientes necesidades de la comunidad oceanográfica brasileña. El elevado número de proyectos puestos en marcha para organizar la gobernanza europea para algunas categorías de disciplinas y redes de organizaciones de investigación marina y grandes iniciativas integradoras (SEADATANET, EMODNET, etc.), han contribuido a reforzar la cooperación entre los consorcios y estándares internacionales. También ha favorecido la mejora de la gestión y la interoperabilidad a escala europea dentro de las infraestructuras de las disciplinas presentadas en esta tesis.

SeaDataNet ha desarrollado un léxico común para datos marinos en todas las disciplinas y aplicaciones y una estructura abierta que puede, con el tiempo, dar acceso a un número cada vez mayor de los centros de datos en todos los sectores y países. El compartimento de los datos podrían proporcionar un marco sólido para el desarrollo estructurado de una red de centros de datos distribuidos utilizando un léxico común para garantizar un

amplio acceso a los usuarios científicos y a los responsables políticos, así como herramientas de fáciles de usar.

Sin embargo, la gestión de los datos marinos exige un marco estratégico para identificar las necesidades fundamentales de los estudios oceanográficos y los objetivos a nivel europeo, y que (no va) prevé un desarrollo coordinado de las diferentes iniciativas, proyectos y repositorios de datos. La consulta actual sobre conocimiento del medio marino en marcha por la Comisión Europea proporciona una oportunidad para desarrollar un marco estratégico para la observación de los océanos en Europa. Tal proceso sería proporcionar una base para un sistema de observación del océano y promover la convergencia entre las diferentes iniciativas europeas, y proporcionar la base para la observación integrada del océano.

Europa deberá mantener una fuerte capacidad de innovación en la observación marina, con el fin de mejorar constantemente la capacidad de supervisar los océanos, posibilitando la mejora de su infraestructura. Un esfuerzo estructurado de investigación a largo plazo debe realizarse en el marco de "Horizonte 2020" en cooperación con otros instrumentos de financiación de la UE (estructurales y los fondos marítimos) para apoyar este objetivo estratégico. Este proyecto está en consonancia con las nuevas políticas de datos de investigación, cada vez más adoptadas por otros organismos de financiación y por las editoriales científicas (como PLoS y Nature Publishing Group).

En relación a Australia, al igual que muchos países en el mundo, ha tenido durante mucho tiempo un centro de datos oceanográficos nacional. Sin embargo, en los últimos años Australia ha cambiado del concepto de un solo centro a una distribución, a modo de red de la (quitar) operación. Así, el gobierno ha creado el Australian Ocean Data Centre Joint Facility (AODC-JF) para proporcionar un enfoque con el control total por parte del gobierno para la gestión de datos del océano australiano. El AODC-JF se estableció bajo una agencia estatal por el gobierno sancionada, con sus Jefes de colaboración de acuerdo. Seis agencias estatales con importantes responsabilidades de datos oceánicos eran signatarios del acuerdo y se comprometieron, en el plano institucional, a hacer considerables tenencias de

datos de libre acceso. El resultado, como hemos visto en el análisis de la situación internacional (cap. 4), es que la ciencia oceanográfica en Australia ahora tiene recursos sumados entre agencias que posibilitan avanzar progresivamente con una eficiente gestión de datos oceanográficos.

Otro país que también se destaca es los Estados Unidos, una vez que el desarrollo de los NODCs actúan de manera integrada con los principales repositorios internacionales, tanto como depositario como en el apoyo para el mantenimiento de las redes de consorcios. Por ejemplo, el NODCs de los Estados Unidos conectan tecnologías submarinas, buques de investigación especializados, observatorios oceanográficos, sistemas de observación de océanos por satélite e *in situ*, instalaciones de supervisión y recogida de datos permanentes, bases de datos y portales de información, medios informáticos de alto rendimiento, instalaciones de modelización y terrestres, unificando la investigación científica del país en un núcleo de datos accesible para todos los centros de investigación y colaborativo entre los consorcios que hemos analizado en el capítulo 4.

En los últimos años, se han tomado muchas iniciativas para el desarrollo de base de datos, permitiendo implementar una arquitectura de información que cumple tanto con la necesidad de la divulgación de los documentos, como con la adecuación necesaria para la interoperabilidad e integración de datos, dirigida a la organización y al uso de la Información científica. Iniciativas fiables y económicas representan la disposición de las interfaces en las que es posible configurar plataformas de acuerdo con las necesidades operacionales de un repositorio, haciendo la recopilación de datos estructurados y sin la necesidad de la implementación de sistemas de su base, o sea, creando nuevos modelos.

Para satisfacer las necesidades de transferencia de datos entre centros de investigación, por medio de los repositorios de cada uno, la interoperabilidad debe seguir los estándares que hemos presentado a lo largo de esta tesis, presentando una estructura fundamental para el intercambio de datos entre repositorios, así como la posibilidad de realización de la búsqueda conjunta de los contenidos compartidos entre los diferentes repositorios.

La norma ISO 19115, adoptada por la mayoría de los repositorios analizados, es un protocolo para la definición y el intercambio de metadatos que tiene cumplido con eficiencia su papel como estándar internacional. La ISO 19115 ha posibilitado que los metadatos puedan ser recogidos por un sistema externo (otro repositorio) para ofrecer un nuevo servicio (por ejemplo, una búsqueda más amplia, el análisis de citas, etc.). A partir de este desarrollo, es posible recoger varios archivos, cambio de registros o buscar disciplinas afines, al mismo tiempo, así como implementar nuevos servicios. En la actualidad, la mayoría de los repositorios utilizan este protocolo para garantizar la integración de sus servicios, lo que permite la difusión y recogida de metadatos, la creación de una red de comunicación entre repositorios interconectados.

La consulta sobre estos repositorios de red, a través de la implementación de la norma ISO 19115 ofrece una única plataforma, con los resultados de la búsqueda de los documentos buscados en repositorios externos, señalando directamente el enlace de descarga o el acceso directo. En consecuencia, la búsqueda simultánea en varios repositorios facilita la consulta de búsqueda de los usuarios para centralizar en una única interfaz, la ampliación de los resultados de búsqueda. Para una institución responsable de mantener un repositorio de libre acceso, debe seguir los protocolos de interoperabilidad que hemos presentado, todavía respetando la capacidad de un servidor, con referencias cruzadas que se encuentran en otros repositorios de metadatos.

Estudio de usuarios y situación en Brasil

Según el estudio de usuarios realizado, un punto de partida para coordinar la adquisición y el acceso a los datos oceanográficos en Brasil reside en el desarrollo de la actual Infraestructura de Datos Espaciales (INDE), empezando por el ámbito nacional, pero con la posibilidad de la globalización, en un esfuerzo conjunto de las instituciones que depositan estos datos.

Después del análisis de las necesidades de los investigadores y de la situación actual de la infraestructura de gestión de datos oceanográficos en

Brasil, se ha puesto de manifiesto que los investigadores brasileños necesitan extraer la información relevante de *datasets* que no están, en su mayoría, integrados en una red común, tampoco estructurados. Esos datos requieren del establecimiento de una arquitectura para gestionarlos y manipularlos, de modo que, para responder dónde deberíamos emprender cuando se trata de extraer valor del entorno de los datos oceanográficos, primero identificamos las ventajas para que los investigadores inviertan en la tecnología necesaria para automatizar el proceso de captura, procesamiento y almacenamiento de datos. Después, planificamos una estrategia de gestión de datos progresiva, que habilite a cada centro de investigación para mantener, entender e interpretar los datos que se manipulan, proporcionando como resultados beneficios tangibles. Con eso presentamos el reto principal para los gestores de información, que es la estandarización y la interoperabilidad entre los repositorios.

Concretamente, los repositorios de datos pueden ser responsables de la custodia de los datos oceanográficos desarrollando estrategias específicas para cada área de estudios. El análisis de las entrevistas demuestra que un enfoque genérico para la preservación de los datos no es suficiente para manejar todas las necesidades y expectativas de los investigadores de diferentes áreas. Es precisamente esta necesidad de combinar la dimensión institucional (muy amplia y multidisciplinaria en el caso de las diferentes investigaciones sobre de los océanos) con la dimensión disciplinaria (con sus requisitos específicos) lo que constituye un desafío importante para el uso de los repositorios de datos oceanográficos como un componente clave en la infraestructura de la preservación general de los datos científicos. Sin embargo, el desafío de fondo es que las investigaciones oceanográficas brasileñas, de un modo general, están extrayendo poco contenido de esos datos, o sea, información realmente eficaz. Eso dicen la encuesta y las entrevistas que hemos realizado: las investigaciones oceanográficas en Brasil con fomento de las universidades y la marina no están integradas, cada una opera de una manera individual.

Para el manejo de datos oceanográficos no existen formatos rígidos, ni un software en particular, lo más importante es seleccionar las herramientas

adecuadas compatibles con los estándares y formatos utilizados por el centro de datos nacional para facilitar la importación y exportación de los conjuntos de datos y metadatos. De esta manera, muchos repositorios se construyeron a través de la reutilización de arquitecturas de información. Tal como los ejemplos citados en el apartado 5.5 (Ibama, ICMBio, Ministerio del medio ambiente, etc), el resultado fue plataformas con costos bajos y confiables. Siguiendo esta necesidad multidimensional del desarrollo de una infraestructura para gestión de datos oceanográficos, consideramos que:

• La mayoría de problemas presentados durante la estandarización obedecen a la falta de comunicación entre las organizaciones productoras de datos del país y a la ausencia de una política nacional de intercambio de datos oceanográficos, que incluya temas tan importantes como los derechos de autor y el control de calidad.

• La actividad de asignar indicadores de calidad debe formar parte de las directrices de la política nacional para ofrecer al usuario información confiable. Estos indicadores deberán asignarse a partir de rigurosas pruebas de calidad que incluyan tanto la documentación (metadatos) de cada una de las etapas y funciones a las que un dato oceanográfico ha sido sometido y, en la medida de lo posible, deben obedecer a procedimientos automatizados.

• Teniendo en cuenta el aporte que brindan los metadatos en el proceso de validación de la calidad de los datos, es importante invertir tiempo en la recopilación de los mismos durante el proceso mismo de obtención de los datos. Este ejercicio puede ahorrarle a las organizaciones productoras dinero y energía que habría que invertir en la realización de arqueología y recuperación; además se compensan los problemas y los costos inherentes a contar con duplicidad, variedad de formatos o redundancia de datos.

• Es necesario desarrollar planes de adquisición de nuevos Sistemas de Información para Brasil; es imperante la activación de un portal web o repositorio de datos que ofrezca la prestación de los servicios de intercambio de datos e información oceanográfica, de acuerdo con el compromiso internacional adquirido con la Comisión Oceanográfica Internacional (COI).

• Acreditación del Banco Nacional de Datos Oceanográficos (BNDO) como un National Oceanographic Data Center (NODC), en respuesta a la nueva estrategia de Gestión de datos y informaciones Oceanográficas adoptada por la Comisión Oceanográfica Intergubernamental (COI) para el Programa Intercambio internacional de Datos e Informaciones Oceanográficas (IODE).

• Desarrollo de un nuevo modelo conceptual una base de datos con estructura para almacenar y recuperar datos y metadatos geoespaciales;

• Optimización de las diversas formas de entrada de datos que se alimentan de la base de datos debido a sus diferentes fuentes y formatos que reciben;

• Creación de un portal de acceso a los datos;

• Optimización del control de recibimiento y disponibilizar los datos;

• Integración del BNDO con el Ocean Data Portal (ODP) del programa IODE de la COI.

Brasil cuenta con dos sectores capacitados para el desarrollo de una infraestructura de gestión de datos oceanográficos: la Infraestructura Nacional de Datos Abiertos (INDA) y la Infraestructura Nacional de Datos Espaciales (INDE).

La INDA "es un conjunto de normas, tecnologías, procedimientos y mecanismos de control necesarios para cumplir con las condiciones de difusión y el intercambio de datos y la información pública sobre el Modelo de Datos Abiertos, de acuerdo con las disposiciones del e-Ping ".

La INDE es el conjunto integrado de tecnologías, políticas y mecanismos de coordinación y procedimientos de supervisión, normas y acuerdos necesarios para facilitar y organizar la producción, el almacenamiento, el acceso, el intercambio, la difusión y el uso de la fuente de datos geoespacial federales, estatales, del condado y municipales. Sin embargo, en la situación brasileña el sistema integrado de observación marina debe esforzarse por proporcionar los flujos de datos necesarios para apoyar un nuevo sistema de gestión de datos oceanográficos que es el enfoque de esta tesis y necesarios para una gestión exitosa en el ámbito marino.

Propuesta de modelo

Para hacer frente a este modelo de búsqueda, nuestra propuesta esta direccionada al modelo Arc Marine debido a la amplitud de sus características que contemplan las demandas en el campo de la oceanografía. Sin embargo, el Arc Marine se puede personalizar para utilizar conforme los requisitos y necesidades que cumplen más adecuadamente a las necesidades específicas de los desarrolladores/administradores de repositorios.

El universo de datos marinos es muy amplio, variado y complejo. Comprende datos paramétricos, atributos de modelización, de infraestructuras, etc, y además es dinámico al contemplar cuatro dimensiones relacionadas con su ubicación (x, y, z, t). Para incorporar datos marinos en el sistema de información geográfica se han empleado las geometrías básicas Common Marine Data Types (Wright *et al*.2000).

El modelo propuesto se ha diseñado para poder incorporar información marina de muy diferente tipologías, generada por distintas organizaciones (centros de investigación, Marina de Brasil, universidades, etc.) y con diferentes instrumentos (metadatos). Cada organización genera los datos en el área que actúan y las actividades se gestionan según los tipos de actividad en función de la naturaleza del dato y se pueden agrupar en proyectos, y éstos en programas.

La idea principal es que el Arc Marine sirva como una plantilla generalizada para guiar la implementación de proyectos de sistemas de información geográfica (GIS) y repositorios de datos del medio marino. Este modelo debe servir para la adaptación de estándares internacionales para facilitar el intercambio de datos y el desarrollo de herramientas analíticas. En este estudio, el Arc Marine se extiende desde su modelo de núcleo para adaptarse a los objetivos de investigación de esta tesis, ofreciendo una alternativa para la gestión de datos de investigación oceanográficos en Brasil.

Limitaciones del estudio

En el desarrollo de la investigación se presentaron las siguientes limitaciones:

La primera limitación se encuentra en la escasez bibliográfica sobre gestión de datos oceanográficos, lo que conllevó extrapolar la información existente sobre el tema en el campo de la Oceanografia y la preservación digital. Debido a que no se cuenta con fuentes de información suficiente que tratan el tema de la gestión de datos oceanográficos, fue necesario ir al Centro Mediterráneo de Investigaciones Marinas y Ambientales - CSIC, ya que una parte fundamental de los conceptos que tratan la gestión de datos marinos requerían mayor comprensión antes desarrollar la encuesta.

La segunda limitación se refiere al estudio de usuarios y específicamente al número de la muestra de investigadores. Aunque se haya aplicado al 100% del universo, si este hubiera sido mayor las posibilidades de encontrar asociación entre la formas como los investigadores gestionan los datos de sus investigaciones para la proponer un modelo global de las necesidades en el caso brasileño. La falta de respuesta de un investigador reduce en una muestra de manera significativa pues el universo de investigadores que trabajan en el campo de la oceanografía brasileña es considerablemente mayor que el numero de respuestas obtenidas en la encuesta, por lo tanto para el constructo se adaptó a los cuarenta y cuatro cuestionarios contestados.

La tercera limitación es la falta de una plataforma de pruebas para el modelo propuesto. Mismo con el entendimiento que el desarrollo de una infraestructura tecnológica que atenta las recomendaciones presentadas en esta tesis sea una etapa *a posteriori,* es decir, que no contempla los objetivos propuestos, sería oportuno aplicar los conceptos propuestos. Eso no fue posible tanto por la falta de recursos económicos y la falta de tiempo necesario cumplir todos los requisitos necesarios.

BIBLIOGRAFIA CONSULTADA

ADAMS, Barbara D.; BRUYN, Lora E.; HOUDE, Sébastien. (2003). Trust In Automated Systems Literature Review. Toronto, ON, Defence R&D Canada.

ALBADALEJO, Pérez; LÓPEZ, J.A. (2011). Sistemas para monitorizar entornos marinos basado en redes sensores inalámbricas. *Jornadas de introducción a la investigación de la UPCT*, Nº. 4, Págs .54-56.

ALIPRANDI, Simone. (2011) Interoperability and open standards: The key to true openness and innovation. *International Free and Open Source Software Law Review* 3: pp. 5–24.

ALOISIO, Giovanni; BARBA, Maria Cristina; BLASI, Euro; CAFARO, Massimo; FIORE, Sandro; MIRTO, Maria. (2004). BIG: A grid portal for biomedical data and images. Journal of Systemics, Cybernetics, and Informatics 2: pp. 10–18 .

ALMEIDA CARNEIRO, A. C. (2001). O uso das estruturas de metadados na implementação da qualidade dos dados de um datawarehouse. Relatório técnico, Universidade Federal do Rio de Janeiro.

ALVES, R. C. V. Web semântica*:* uma análise focada no uso de metadados. (2005). 180 f. Dissertação (Mestrado em Ciência da Informação) - Faculdade de Filosofia e Ciências, Universidade Estadual Paulista, Marília.

AMERICAN GEOPHYSICAL UNION. AGU publications data policy. (2013). http://publications.agu.org/author-resource-center/publication-policies/data-policy/

ANDREWS, B. (2006). Managing seafloor mapping data using the Arc Marine data model. In Proceedings of the Geographic Information Systems and Ocean Mapping in Support of Fisheries Research and Management, Cambridge, Massachusetts: pp. 30–33

ANTARCTIC MASTER DIRECTORY. http://gcmd.gsfc.nasa.gov/KeywordSearch/amd/about_us.html

AOYAMA M.; HIROSE, K. (2004). Artificial radionuclides database in the Pacific Ocean: HAM database. The Scientific World 4: pp. 200–215.

ARZBERGER, P.; SCHROEDER, P.; BEAULIEU, A.; BOWKER, G.; CASEY, K.; LAAKSONEN, L.; MOORMAN, D.; UHLIR, P.; WOUTERS, P. (2004) Promoting access to public research data for scientific, economic, and social development. *Data Science Journal*, 3, pp. 135-152.

AUSTRALIAN Ocean Data Centre (2008). Marine Community Profile of ISO 19115. Canberra, ACT, Australian Ocean Data Centre.

BACA, Murtha. (2008). Introduction to metadata. Los Angeles: Getty research institute.

BAKER, R. (1990) CASE. Method: Entity Relationship Modelling. Cambridge, MA, Addison-Wesley

BAKER, K. S.; CHANDLER, C. L. (2008). Enabling long-term oceanographic research: Changing data practices, information management strategies and informatics. Deep Sea Research II 55: pp. 2132–2142

BALLAGH, L. M.; RAUP, B. H.; DUERR, R. E.; KHALSA, S. J. S.; HELM, C.; FOWLER, D.; GUPTE, A. (2011). Representing scientific data sets in KML: Methods and challenges. *Computers and Geosciences* 37: pp. 57–64

BANCO de Datos de Datos Ambientales para la Industria Petrolera (Banpetro). (2016). http://www.bampetro.on.br

BANCO NACIONAL DE DATOS OCEANOGRÁFICOS. 2016. https://www.mar.mil.br/dhn/chm/bndo/bndoeoiode.htm

BARREIRO, Wulff Enrique. (2011). Aproaches to open data for science in Spain. Data Science Journal, v. 10, n. 27. https://www.jstage.jst.go.jp/article/dsj/10/0/10_10-001/_article

BEARE, D.; KENNY, A.; KERSHAW, P.; MCKENZIE, E.; DEVLIN, M.; REID, J.; LICANDRO, P.; WINPENNY, K.; HAUGHTON, C.; LANGSTON, M.; SKJOLDAL, H. R.; PERKINS, A. (2006). Building multi-discipline, multivariate databases for use in integrated ecosystem assessments: Experiences and recommendations. In *Proceedings of the ICES 2006 Annual Science Conference*, Maastricht, The Netherlands

BECHINI, A.; VETRANO, A. (2013) Management and storage of in situ oceanographic data: An ECM-based approach. *Information Systems* 38: 351–368.

BENAVENT, Aleixandre; INFER, Antonio Vidal; ARROYO, Adolfo Alonso; SA-PENA, A. Ferrer, PESET, Fernanda; GARCÍA, A. García. (2014). Gestión de los datos brutos de investigación en los investigadores españoles en ciencias de la salud. Trauma Fund. Mapfre. Vol. 25, n. 4.

BERMUDEZ, O.; BARRAGÁN, A.; ALONSO, F. (2011). La gestión de los datos polares em España: una aproximación a la contribución de las ciencias de la vida. Ecosistemas, v. 20, n. 1, pp. 94-103, Jan.

BJÖRK, B. (2005). A lifecycle model of the scientific communication process. Learned publishing. v. 18, n. 3, p. 165-176. http://oacs.shh.fi/publications/model35explanation2.pdf

BLANC, F., CLANCY, R., Comillon, P., Donlon, C., Hacker, P., Haines, K., Hankin, S., Pouliquen, S., Price, M., Pugh, T. & Srinivasan, A. (2008). Data &

product serving, an overview of capabilities developed in 10 years. *GODAE Final Symposium*, Nice, France.

BLOWER, J. D., BLANC, F., CLANCY, R. M., CORNILLON, P., DONLON, C., HACKER, P., HAINES, K., HANKIN, S. C., LOUBRIEU, T., POULIQUEN, S., PRICE, M., PUGH, T. F.; SRINIVASAN, A. (2009). Serving GODAE data and products to the Ocean Community. *Oceanography*, vol. 22.

BORGMAN, Christine L. (2007). Scholarship in the digital age: information, infraestructure, and the internet. Cambridge: MIT Press.

BORGMAN, Cristine L. (2010). Research data: who will share what, with whom, when, and why? In: CHINA NORTH AMERICAN LIBRARY CONFERENCE, 5., Beijing. http://works.bepress.com/borgman/238/

BORGMAN, Christine L. (2012). The conundrum of sharing research data. *Journal of the American Society for Information Science and Technology .* n. 63, pp. 1059–1078.

BORREGO, Àngel (2012). Los retos de la gestión de datos de investigación. *Blok de bid*, 6 nov. http://www.ub.edu/blokdebid/es/content/los-retos-de-la-gestión-de-datos-de-investigación

BOYER, T. P.; ANTONOV, J.; GARCIA, H. E., *et al*. (2006) *World Ocean Database* 2005 (ed. Sydney Levitus). NOAA Atlas NEDIS 60. Washington, DC: US Government Printing Office. DVD, 190 pp.

BRASE, J. (2004). *Using digital library techniques* - Registration of scientific primary data. In: HEERY, R; LYON, L (org). Research and Advanced Technology for Digital Libraries. pp.. 488–494.

BRASIL. Presidência da República. (2008). Decreto n° 6.666, de 27 de novembro de 2008. Institui a Infraestrutura Nacional de Dados Espaciais - INDE. Diário Oficial da República Federativa do Brasil, Braslília.

BRITO, T. Antártida Bem comum da Humanidade. Brasília: Ministério do Meio Ambiente.

BUITENHUIS, E.T.; VOGT, M.; MORIARTY, R.; BEDNARŠEK, N.; DONEY, S. C.; LEBLANC, K.; LE QUÉRÉ, C.; LUO, Y.-W.; O'BRIEN, C.; O'BRIEN, T.; PELOQUIN, J.; SCHIEBEL, R.; SWAN, C. (2013). MAREDAT: towards a world atlas of MARine Ecosystem DATa. *Earth Syst. Sci. Data*, 5, pp. 227-239.

CARLSON, Jake R. (2012). Demystifying the Data Interview: Developing a Foundation for Reference Librarians to Talk with Researchers about their Data. *Reference Services Review* v. 40, n. 1. p. 7-23.

CASTILLO, José Manuel Morales del (2011). Hacia la biblioteca digital semántica. Gijón: Trea.

CASTRO, Fábio de. (2012). Falta de uma infraestrutura de dados espaciais limita pesquisa oceanográfica no Brasil, diz especialista. FAPESP. http://agencia.fapesp.br/15472

CATERIANO, Edgar. (2010). Gestionando con conocimiento: La inteligencia al servicio de las organizaciones. http://migre.me/qEXvM

CENTRO NACIONAL DE DATOS POLARES, Instituto Geológico y Minero de España. (2004). Protocolo de remisión, almacenamiento y difusión de los datos antárticos. http://hielo.igme.es

CERN. http://home.web.cern.ch

CGEE. Avaliação preliminar do Programa Antártico Brasileiro Brasília. (2006). http://www.cgee.org.br/atividades/redirect/3422

CHOWDHURY, Gobinda G.; FOO, Schubert. (2012). Digital Libraries and information Acess: research perspectives. Grã Bretanha: Facet publishing.

CINKOSKY, M. J.; FICKETT, J. W.; GILNA, P.; BURKS. C. (1991). Electronic data publishing and genbank. Science. n. 252. pp. 1273–1277.

CIRANO, M.; MATA, M. M.; CAMPOS, E. J.; DEIRÓ, N. F. (2006). A circulação oceânica de larga-escala na região oeste do atlântico sul com base no modelo de circulação global occam. *Revista Brasileira de Geofísica*, v. 24, n.2, pp. 209–230.

CLOCKSS. http://www.clockss.org/clockss/Home

CNPq Proantar. http://www.cnpq.br/programas/proantar/index.htm

COATES, H.; KONKIEL, S.; WITT, M. (2013). Data Services: Making It Happen. https://scholarworks.iupui.edu/handle/1805/3278

COCHRANE, Jack. (2013). Open by Default: making Open Data truly open. *BioMed Central Blog*. http://blogs.biomedcentral.com/bmcblog/2013/08/21/open-by-default-making-open-data-truly-open/

COCHRANE, N. A. (2007). *Ocean Bottom Acoustic Observations in the Scotian Shelf Gully During an Exploration Seismic Survey: A Detailed Study*. Dartmouth, NS, Canadian Technical Report of Fisheries and Aquatic Sciences N. 2747.

CODATA-ICSTI Data Citation Standards and Practices. (2013). Out of Cite, Out of Mind: The Current State of Practice, Policy, and Technology for the Citation of Data. *Data Science Journal*. Vol. 12.

COMISIÓN OCEANOGRÁFICA INTERGUBERNAMENTAL DE LA UNESCO. (2007). *Plan estratégico de la COI para la gestión de datos e información oceanográficos* (2008–2011). 24a reunión de la asamblea. UNESCO, París, 19–28 de jun.

COMITÊ EXECUTIVO DE GOVERNO ELETRÔNICO. (2004). *e*-PING padrões de interoperabilidade de governo eletrônico – documento de referência versão 0: parte II – especificação dos componentes da *e*-PING. [S.l.]. 64 p. http://www.governoeletronico.gov.br/governoeletronico/publicacao/down_anexo.wsp?tmp.arquivo=E15_24115_1e-ping_minuta_v0_31052004_consulta.pdf

CONICYT, IDER. (2010). Estado del Arte Nacional e Internacional en materia de gestión de datos científicos e Información Científica y Tecnológica y Recomendaciones de Buena Prácticas. Santiago: Gobierno de Chile.

CONKRIGHT, M.E.; LEVITUS, Sydney. (1996) Objective analysis of surface chlorophyll data in the northern hemisphere. In: *Proceedings of the International Workshop on Oceanographic Biological and Chemical Data Management*. NOAA Technical Report NESDIS 87, pp. 33–43.

CONKRIGHT, M. E.;. LEVITUS, SYDNEY. (1996). Objective analysis of surface chlorophyll data in the northern hemi- sphere. In: *Proceedings Of The International Workshop On Oceanographic Biological And Chemical Data Management*. NOAA Technical Report NESDIS, 87, pp. 33-43.

CONSEJO SUPERIOR DE INVESTIGACIONES CIENTÍFICAS. (2016). http://www.csic.es/*web*/guest/home

CONTI, Luis Américo; Oliveira, Mariana Cabral de; ESTRADA, Tiago Egger Moellwald Duque; MARQUES, Antônio Carlos. (2013). Gerenciamento de dados marinhos no contexto brasileiro. Biota Neotropica. Campinas, v. 13, n. 2, pp. 21-26, jun.

COOPER, D. R.; SCHINDLER, P. S. (2003). *Business research methods*. New York: McGraw-Hill.

CORTI, Louise, et al. (2011). *Managing and Sharing Data*: a guide to good practice. Sage Publications.

COSTA, S. M. S. (1999). The impact of computer usage on scholarly communication amongst academic social scientists. 302 p. Tese (Doutorado em ciência da informação) - Loughborough University, Department of Information Science, Loughborough, Inglaterra.

CREATIVE Commons. (2015). creativecommons.org

DAHER, E.; BRITO, T. (Coord.). (2007). O Brasil e o meio ambiente antártico: ensino fundamental e médio. Brasília: Ministério da Educação.

DATA ARCHIVING AND NETWORKED SERVICES. (2016). Data Management Plan for scientific research. http://www.dans.knaw.nl/sites/default/files/file/Datamanagementplan%20UK(1).pdf.

Data Management Plans. (2016). Digital Curation Centre. https://dmponline.dcc.ac.uk

DATABIB. (2016). www.databib.org

DATACITE. (2016). https://www.datacite.org

DEMPSEY, Lorcan; HEERY, Rachel. (1997). A review of metadata: a survey of current resource description formats. Mar. http://www.ukoln.ac.uk/metadata/desire/overview/overview.pdf

DIGITAL Curation Centre. (2016). DCC Curation Lifecycle Model.http://www.dcc.ac.uk/resources/curation-lifecycle-model

Digital Equipment Corporation (ed). (1992). Information Technology – Database Language SQL. Geneva, Switzerland, International Organization for Standardization.

DITTERT, Nicolas; Diepenbroek, Michael; GROBE, Hannes. (2001). Scientific data must be made available to all. *Nature*, v. 14, n. 393.

DIRECTORY of Open Acess Journals. (2016). http://www.doaj.org

DMP TOOL. (2016). https://dmp.cdlib.org

DODGE, Chris; MAJEWSKI, Frank.; MARX, Beate; PFEIFFENBERGER, H; et. al. (1996). Providing global access to marine data via the World Wide Web. *Journal of Visualization and Computer Animation.* n. 7. pp. 159–168.

DRYAD DIGITAL REPOSITORY. (2015). http://datadryad.org

DUBLIN CORE ANNUAL CONFERENCE (2005). Proceedings of the international conference on Dublin Core and Metadata applications. Madrid: Universidad Carlos III de Madrid. Set.

ELFEKY, M. G. y VERYKIOS, V. S. (2002). A record linkage toolbox. International Conference on Data Engineering (ICDE), pp. 17-28. http://www.cs.purdue.edu/homes/mgelfeky/Papers/icde02.pdf

EMODNET. (2016). European Marine Observation and Data. http://bio.emodnet.eu/

ENGINEERING and Physical Sciences Research Council. Expectations. (2014). https://www.epsrc.ac.uk/about/standards/researchdata/expectations/
EUROPEAN Comission. (2007). Scientific information in the digital age: Ensuring current and future access for research and innovation. Brussels. Feb. http://europa.eu/rapid/press-release_IP-07-190_en.htm?locale=en

EUROPEAN Comission. (2012). Towards better access to scientific information: Boosting the benefits of public investment in research. http://

ec.europa.eu/research/science-society/document_library/pdf_06/era-communication-towards-better-access-to-scientific-information_en.pdf

European Commission. *Scientific data*: open access to research results will boost Europe's innovation capacity. (2012). httpeuropaeurapidpress-ReleaseIP-12–790enhtm

European Commission. (2013). *Guidelines on Open Access to Scientific Publications and Research Data in Horizon 2020*. v. 1.0. Dec.

FACHIN, G. et al. (2009). Gestão do conhecimento e a visão cognitiva dos repositórios institucionais. *Perspect. ciênc. inf.*, Belo Horizonte, v. 14, n. 2.

FALCONE, Andres Araya. *Big Data:* El 90% de los datos existentes en el mundo han sido creados en los últimos 2 años. (2011). http://andresarayafalcone.blogspot.com.es/2011/10/big-data-el-90-de-los-datos-existentes.html

FLETCHER M; CLEARY J; COTHRAN J; PORTER D. (2008). Southeast Atlantic Coastal Ocean Observation System (SEACOOS) information management: Evolution of a distributed community system. *Marine Technology Society Journal* 42: pp. 28–34

FORCE11. (2014). J*oint declaration of data citation principles*. https://www.force11.org/group/joint-declaration-data-citation-principles-final

FOULONNEAU, Muriel; RILEY, Jenn. (2008). Metadata for digital resources: implementation, systems design and interoperability. London: Chandos.

FOWLER, M.; SCOTT, K. (1999). UML Distilled: A brief Guide to the Standard Object Modeling Language (Second Edition). Harlow, Addison Wesley Longman.

FROESE, R.; LLORIS, D. Opitz (2003). The need to make scientific data publicly available – concerns and possible solutions. In: *Fish Biodiversity: Local Studies as Basis for Global Inferences.* M.L.D. Palomares, B. Samb, T. Diouf, et al. (Org.), pp. 267–271. Brussels.

FROESE, R. Pauly, D. (eds.). (2009). FishBase. World Wide Web electronic publication. www.fishbase.org

GONZALEZ, M. (2010). Análise das restrições de acesso a dados de espécies ameaçadas, previstas em políticas de coleções biológicas científicas brasileiras, à luz do direito ambiental e da ciência da informação. *Ciência da Informação*, Brasília, v. 39, n. 1, abr.

Fundación para el Conocimiento madri+d. (2016). www.madrimasd.org/informacionidi/e-ciencia/default.asp

GARRITANO, Jeremy R.; CARLSON, Jake R. (2009). A Subject Librarian's guide to Collaborating on e-Science Projects. *Issues in Science and Technology Librarianship*, n. 57.

GONZÁLEZ, Luis-Millán; SAORÍN, Tomás; FERRER-SAPENA, Antonia; ALEIXANDRE-BENAVENT, Rafael; PESET, Fernanda. (2013). Gestión de datos de investigación: infraestructura para su difusión. *El profesional de la información*, septiembre-octubre, v. 22, n. 5

GOOGLE FLU TRENDS. (2016). Conoce la evolución de la gripe en todo el mundo. https://www.google.org/flutrends

GORALSKI R. I.; Gold C. M. (2007). The development of a dynamic GIS for maritime navigation safety. In *Proceedings of the ISPRS Workshop on Updating Geo-spatial Databases with Imagery and the Fifth ISPRS Workshop on DMGISs*, Urumchi, China: 47–50.

GORGOLEWSKI, Krzysztof; MARGULIES, Daniel S.; MILHAM, Michael P. (2013). Making data sharing count: a publication-based solution. *Frontiers in Neuroscience*. v. 7, n. 9.

GRAAF, Maurits van der; WAAIJERS, Leo. (2011). *The Riding the Wave report*. http://www.knowledge-exchange.info/default.aspx?id=469

GRAYBEAL, J.; ISENOR, A. W.; REUDA, C. (2012). Semantic mediation of vocabularies for ocean observing systems. *Computers and Geosciences* 40: pp. 120–131.

Grupo de Trabajo de “Depósito y Gestión de datos en Acceso Abierto” del proyecto RECOLECTA. (2012). *La conservación y reutilización de los datos científicos en España. Informe del grupo de trabajo de buenas prácticas.* Madrid: Fundación Española para la Ciencia y la Tecnología, FECYT. http://www.recolecta.net/buscador/documentos/informe_datos_cientificos_en_esp.p

HARNAD, Stevan. et al. (2004). The Access/Impact Problem and the Green and Gold Roads to Open Access. *Serials review*. v.30, n. 4. http://eprints.soton.ac.uk/260209

Harvard University. Retention of research data and materials. (2011). Harvard University Office of Sponsored Programs. http://osp.finance.harvard.edu/retention-research-data-and-materials

HARVEY, Ross. (2010). *Digital Curation*: a how-to-do-it manual. New York: Neal-Schuman.
HAY, David C. (2006). *Data model patterns*: a metadata map. San Francisco: Elsevier.

HAYNES, David. (2004). Metadata: for information management and retrieval. London: Facet.

HEAVNER, M. J.; FATLAND, D. R.; HOOD, E.; CONNER. C. (2011). SEAMONSTER: A demonstration sensor web operating in virtual globes. *Computers and Geosciences* 37: pp. 93–99.

HERNÁNDEZ-JAIMES, José Luiz. (2008). Gestión de datos e información oceanográfica Colombiana. Centro Control Contaminación del Pacífico. Serie de Publicaciones Especiales, v. 6, San Andrés de Tumaco.

HERNÁNDEZ-PÉREZ, Tony; GARCÍA-MORENO, María-Antonia. Datos abiertos y repositorios de datos: nuevo reto para los bibliotecarios. *El profesional de la informacion*, 2013, vol. 22, n. 3, pp. 259-263.

HEY, Tony; TANSLEY, Stewart; TOLLE, Kristin, eds. (2009). *The Fourth Paradigm*: Data-Intensive Scientific Discovery. Redmond,Washington, Microsoft Research.

HIGGINS, Sarah. (2007). The DCC curation lifecycle model. *Proceedings of the 8th ACM/IEEE-CS joint conference on Digital libraries*. Pittsburgh PA, PA, USA, pp. 453-453.

HILLMANN, Diane I; WESTBROOKS, Elaine L. (2004). *Metadata in practice*. Chicago: ALA editions.

HILLMANN, Diane I. (2013). Table of contents. http://dublincore.org/documents/2001/04/12/usageguide.

HOEVEN, Jeffrey van der. (2010). Insight into digital preservation of research output in Europe.. http://www.parse-insight.eu/downloads/PARSE-Insight_D3-6_InsightReport.pdf

HOLDREN, John P. (2013). Increasing Access to the Results of Federally Funded Scientific Research. Office of science and technology policy.

HOUYOUX, M.; STRUM M.; MASON, R. (2006). Data management using the emissions modeling framework. In *Proceedings of the Fifteenth International Emission Inventory Conference*, New Orleans, Louisiana

HUTT, D.; OSLER, J.; ELLIS, D. (2002). Effect of Hurricane Michael on the underwater acoustic environment of the Scotian Shelf. In *Proceedings of the Impact of Littoral Environmental Variability of Acoustic Predictions and Sonar Performance*, Lerici, Italy: pp. 27–34.

ICAN (Internacional Coastal Atlas Network. (2016). http://ican.science.oregonstate.edu

INSPIRE (Infraestructura de información espacial en la Comunidad Europea. (2016). http://inspire.jrc.ec.europa.eu

INSPIRE. (2010). INSPIRE Metadata Implementing Rules: Technical Guidelines based on EN ISO 19115 and EN ISO 19119. Ispra, Italy, European Commission Joint Research Centre.

INTERGOVERNMENTAL Oceanographic Commission. Manuals and Guides. http://goo.gl/wD9lWQ

IODE (International Oceanographic Data and Information Exchange). (2016). http://www.iode.org/

ISENOR, Anthony W.; STUART, Robert A. (2007). The utilization of web services while at sea. *CMOS Bulletin* 35: pp. 17–19

ISO 2003. Geographic Information – Metadata. Geneva, Switzerland, International Organization for Standardization.

KIM, Young-Gul; MARCH Salvatore T. (1995). Comparing data modeling formalisms. *Communications of the ACM* 38: 103–113.

KNOWLEDGE EXCHANGE. (2016). http://www.knowledge-exchange.info/

KOTARSI, Rachael Kotarski; REILLY, Susan Reilly; SCHRIMPF, Sabine; SMIT, Eefke; WALSHE, Karen. (2012). Report on best practices for citability of data and on evolving roles in scholarly communication.

KOKKONEN, T, JOLMA, A; KOIVUSALO H. (2003) Interfacing environmental simulation models and databases using XML. *Environmental Modelling and Software* 18: 463–471

KRATZ, John. (2013). Data citation developments. Data Pub. http://datapub.cdlib.org/2013/10/11/data-citation-developments/.

KUPCA, V. (2004). A standardized database for fisheries data. In *Proceedings of the ICES 2004 Annual Science Conference*. Vigo, Spain.

LAGE, Kathryn; LOSOFF, Barbara; MANESS, Jack. (2011). Receptivity to Library Involvement in Scientific Data Curation: A Case Study at the University of Colorado Boulder. *Libraries and the Academy*, v. 11, N. 4. pp. 915–937.

LAWRENCE, S. (2001). Free online availability substantially increases a paper's impact. *Nature*. http://www.nature.com/nature/debates/e-access/Articles/lawrence.html

LEITE, Fernando Cesar Lima. (2009). *Como gerenciar e ampliar a visibilidade da informação científica brasileira*: repositórios institucionais de acesso aberto. Brasília: IBICT.

LEITE, F.; COSTA, S. (2006). Repositórios institucionais como ferramentas de gestão do conhecimento científico no ambiente acadêmico. *Perspect. ciênc. inf.*, Belo Horizonte, v. 11, n. 2, ago.

LERU Roadmap for research data Advice paper, n. 14. Dic. 2013. http://goo.gl/xvyrjS

LEUNG, Felix.; BOLLOJU, Narasimha. (2005). Analyzing the quality of domain models developed by novice systems analysts. In *Proceedings of the Thirty-eighth Hawaii International Conference on System Sciences*, Big Island, Hawaii: 188

LEVITUS, Sydney; ANTONOV, J. I.; BARANOVA, O. K.; BOYER, T. P.; COLEMAN, C. L.; GARCIA, H. E.; GRODSKY, a I.; JOHNSON, D. R.; LOCARNINI, R. A.; MISHONOV, A. V., et al. (2013). The World Ocean Database. *Data Science Journal*, 12 maio.

LEVITUS, Sydney. Interannual-to-decadal variability of the temperature-salinity structure of the world ocean. In: *Proceedings of the International Workshop on Oceanographic Biological and Chemical Data Management*. NOAA Technical Report NESDIS, 87, pp. 51-54. 1996.

LEVITUS, Sydney. The UNESCO-COI-IODE Global Oceanographic Data Archeology and Rescue (GODAR) Project and World Ocean Database"projects. *Data Science Journal*, 11, pp. 46–71. 2012.

LIBER 40th Annual Conference. (2011). Universitat Politècnica de Catalunya. Barcelona, 29 June. Julyhttp://liber2011.upc.edu

LIU, Jia. (2007). *Metadata and its applications in the digital library*: approaches and practices. London: Publishing Group.

LORD-CASTILLO, B. K; WRIGHT, D. J.; MATE, B. R.; FOLLETT, T. (2009). A customization of the Arc Marine data model to support whale tracking via satellite telemetry. *Transactions in GIS* 13: pp. 63–83

MACHADO, M.; BRITO, T. (2006). Antártida: ensino fundamental e ensino médio. Brasília: Ministério da Educação.

MACHADO, Antonio.; BERMEJO, J. A. *Proyecto REDMIC*. (2010). Repositorio de Datos Marinos Integrados de Canarias. OAG, Observatorio Ambiental Granadilla Santa Cruz de Tenerife.

MANYIKA, et al. (2011). *Big data*: The next frontier for innovation, competition, and productivity. McKinsey Company. http://goo.gl/RQdzh

MANUAL de Periodismo de Datos. (2016). http://interactivos.lanacion.com.ar/manual-data

MARINHA DO BRASIL. (2016). http://www.mar.mil.br
Marine Metadata Interoperatibility Project. http://marinemetadata.org/

MCCANN, Michael; GOMES, Kevin (2008). *Oceanographic data provenance tracking with the shore side data system*. In Freire J , Koop D , and Moreau L (eds) IPAW 2008. Berlin, Springer-Verlag Lecture Notes in Computer Science Vol. 5272: pp. 309–322

Medical Research Council. Section 2: Guidelines and Standards. (2014). http://www.mrc.ac.uk/documents/pdf/good-research-practice-guidelines-and-standards/

MEGHINI, C. (2013). Data preservation. *Data Science Journal*, v. 12. jul.

MELERO, Remedios; HERNÁNDEZ-SAN-MIGUEL, Javier. Acceso abierto a los datos de investigación, una vía hacia la colaboración científica. *Revista española de Documentación Científica*, v. 37, n. 4. 2014.

MELERO, Remedios. (2010). Una pleamar de datos. *Reseñas de Biblioteconomía y Documentación.* http://www.ub.edu/BLOKDEBID/ES/CONTENT/UNA-PLEAMAR-DE-DATOS

MÉNDEZ RODRÍGUEZ, Eva. (2002). Metadatos y recuperación de la información: estándares, problemas y aplicabilidad en bibliotecas digitales. Gijón, Trea.

MERCEUR, F. Référencement et Attribution de DOI à des jeux de données géographiques *Journée Administrateurs Sextant*, France. 2014. Disponible en: < https://www.ifremer.fr/sextant_doc/sextant/sextant/jas_2014/jas_2014_doi.pdf >. Acceso en: 13 dic. 2014.

MESH (Mapping European Seabed Habitats). (2016). http://www.searchmesh.net

MGDA (Marine Geophysical Data Access;. Disponible en: < http: // www.ngdc.noaa.gov/mgg/geodas). Acceso en: 31 ene. 2016.

MOODY, D. L.; SHANKS, G. G. (1994). What makes a good data model?: Evaluating the quality of entity-relationship models. In *Proceedings of the Thirteenth International Conference on the Entity-Relationship Approach*, Manchester, United Kingdom: pp. 94–111

MOODY, D. L.; SHANKS, G. G. (2003). Improving the quality of data models: Empirical validation of a quality management framework. *Information Systems* 28: pp. 619–650

MOODY, D. L.; SINDRE, G., BRASETHVIK, T.; SOLVBERG A. (2003). Evaluating the quality of information models: Empirical testing of a conceptual model quality framework. In *Proceedings of the Twenty-fifth International Conference on Software Engineering*, Portland, Oregon: pp. 295–305.

MOONEY, Hailey; NEWTON, Mark P. The anatomy of a data citation: discovery, reuse, and credit. *Journal of a Librarianship and Scholarly Communication*, vol. 1, n.1, pp. 1035.

MORENO, Fernanda Passini; LEITE, Fernando César Lima; ARELLANO, Miguel Ángel Márdero. Acesso livre a publicações e repositórios digitais em ciência da informação no Brasil. Perspect. ciênc. inf., Belo Horizonte, v. 11, n. 1, abr. 2006.

MOSSINK, Wilma; BIJSTERBOSCH, Magchiel; NORTIER, Joeri. (2013). SIM4RDM. *European Landscape Study of Research Data Management*. http://www.sim4rdm.eu

NASA. Directory interchange format (DIF) Writers guide. Global change master directory. National Aeronautics and Space Administration. http://gcmd.gsfc.nasa.gov/difguide/difman.html

NATIONAL DATA BUOY CENTER. (2016). http://www.ndbc.noaa.gov

NATIONAL Health and Medical Research Council; Australian Government; Australian Research Council. Australian code for the responsible conduct of research. Revision of the joint nhmrc/avcc statement and guidelines on research practice. (2007). http://www.nhmrc.gov.au/_files_nhmrc/publications/attachments/r39.pdf

NATIONAL SCIENCE BOARD. *Long lived digital data collections*: Enabling research and education in the 21st century. http://www.nsf.gov/pubs/2005/nsb0540

NATIONAL SCIENCE FOUNDATION CYBERINFRASTRUCTURE COUNCIL. Cyberinfrastructure vision for 21st century discovery. (2007). http://www.nsf.gov/pubs/2007/nsf0728/nsf0728.pdf

NATIONAL SCIENCE FOUNDATION. Dissemination and Sharing of Research Results. (2015). http://www.nsf.gov/bfa/dias/policy/dmp.jsp

NDE/Brasil (Infraestructura Nacional de Datos Espaciales). (2016). http://www.inde.gov.br

NETWORK Marine Research Institutes and Documents - MARENET. (2016). http://www.marenet.de/MareNet/

NOAA. ETOPO2v2. (2010) ETOPO2v2 Global Gridded 2-minute Database. Washington, D.C., National Geophysical Data Center, National Oceanic and Atmospheric Administration, U.S. Department of Commerce. http://www.ngdc.noaa.gov/mgg/global/etopo2.html

NOAA National Marine Data Center. (2016). http://www.lib.noaa.gov/index.html

NOAA. (2008). National Oceanic and Atmospheric Administration. Procedure for Scientific records appraisal and archive approval: Guide for Data Managers. https://www.ngdc.noaa.gov/wiki/images/0/0b/NOAA_Procedure_document_final.pdf >. Acceso en: 13 ene. 2016

NOAKES, T. D. The limits of endurance exercise. Basic Research in Cardiology, 101(5), 408-417, 2006.

OBIS (Ocean Biogeographic Information System. (2016). http: //www.iobis.org/

ODISEA. International Registry of Research Data. (2016). http:// odisea.ciepi.org

OECD. (2007). OECD: Declaration on Access to Research Data From Public Funding. Paris: OECD. http://www.oecd.org/document/15/0,3343,en_2649_34487_25998799_1_1_1_1,00.htm

OECD. (2004). Declaration on Access to Research Data From Public Funding, Paris. http://goo.gl/lovbt7

OECD Principles and Guidelines for Access to Research Data from Public Funding. (2007). http://www.oecd.org/sti/sci-tech/38500813.pdf

OFFICE of Science and Technology Policy. (2015). https://goo.gl/xZ9zDz >.

OGC Open Geospatial Consortium. (2016). http://www.opengeospatial.org/

OLAYA, Víctor. Sistemas de información geográfica. (2016). http:// volaya.github.io/libro-sig/index.html

ONOFRE, M. (2001). Rapid Environmental Assessment. Anais do Instituto Hidrográfico. n. 15, pp. 25-31.

OPEN AIRE. (2016). https://www.openaire.eu/

OPEN DATA COMMONS: legal tools for Open Data. (2016). opendatacommons.org/licenses

OPEN DATA COMMONS. Public Domain Dedication and License (PDDL). (2008). http://opendatacommons.org/licenses/pddl/1-0/.

OPEN DATA COMMONS: legal tools for Open Data. (2016). opendatacommons.org/licenses

OPEN DATA COMMONS. (2008). Public Domain Dedication and License (PDDL). http://opendatacommons.org/licenses/pddl/1-0/

OPENDOAR. (2016). Directory of open acess repositories. www.opendoar.org

OPEN knowledge foundation. *Open data*: An introduction. http://okfn.org/opendata/

OPEN Knowledge Foundation. *What is Open Data?*. http://okfn.org/opendata

OPPORTUNITY FOR DATA EXCHANGE. (2012). Report on best practices for citability of data and on evolving roles in scholarly communication. http://goo.gl/fbNl4u

ORTIZ-MARTINEZ, R. V. , MOGOLLÓN DIAZ, A. Y. RICO-LUGO, H. D. (2008). Implementation of international standard of Colombian oceanographic data anda information management using open source web software. Case study. International conference on marine data and information systems-IMDIS2008. Atenas, Grecia.

ORTIZ-MARTINEZ, R. V., RODRIGUEZ-RUBIO, E. (2007). Arquitectura base para el intercambio de datos oceanográficos colombianos. Boletín Científico, n.14. Centro Control Contaminación del pacífico. Tumaco: Colombia.

ORTIZ-MARTÍNEZ, R. Introducción a la gestión de datos oceanográficos. (2008). In: Dimar (Ed.). *Gestión datos e información oceanográfica colombiana*. (pp. 43-62). Bogotá: Editorial Dimar.

OSLER John; FURLONG Arnold; CHRISTIAN Harold; LAMPLUGH M. (2006). The integration of the free fall cone penetrometer (FFCPT) with the moving vessel profiler (MVP) for the rapid assessment of seabed characteristics. *International Hydrographic Review* 7(3): pp. 45–54.

PACHECO, B., Martinho, S. (2005). Apoio Ambiental ao Exercício Lusíada 2006. Produtos e Inovações. *Anais do Instituto Hidrográfico*, N.18, pp. 85-94.

PAINEL LATTES. (2016). Conselho Nacional de Desenvolvimento Científico e Tecnológico. http://estatico.cnpq.br/painelLattes/mapa

PALINKAS, L.; SUEDFELD, P. Psychological effects of polar expeditions. *The Lancet*, *371*(9607), pp. 153-163, 2008.

PARSONS Mark A.; GODOY, Oystein; LeDrew Ellsworth; et. al. (2011). A conceptual framework for managing very diverse data for complex, interdisciplinary science. *Journal of Information Science* 37: pp. 555–569.

PATERSON; G., Boxshall, G.; Thomson, N.; Hussey, C. (2000) Where are all the data? *Oceanography* 13(3), pp. 21–24.

PECCOL, Elisabetta. (2004). An ISO 19115 discovery metadata profile for spatial and non-spatial environmental datasets. In *Proceedings of the Tenth EC-GI GIS Workshop*, Warsaw, Poland.

PECKNOLD, Sean; OSLER, John C. (2012). Sensitivity of acoustic propagation to uncertainties in the marine environment as characterized by

various rapid environmental assessment methods. *Ocean Dynamics* 62: pp. 265–281.

PETERS, Christie; DRYDEN, Anita Riley. Assessing the academic library's role in campus-wide research data management: a first step at the University of Houston. *Science and Technology Libraries*, v. 30. n.4. pp. 387-403. 2011.

PETERSON, Jane; CAMPBELL, Joseph. (2010). Marker papers and data citation. *Nature Genetics.* n. 42.

POLOCZANSKA, E., Hobday, A.J.; RICHARDSON, A.J. Global database is needed to support adaptation science. *Nature* 453, 720. 2008.

QUANTUM GIS Development Team. (2011). Quantum GIS. http://www.qgis.org/

RABIANSKI, Joseph. Primary and secondary data: concepts, concerns, errors and issues. *The Appraisal Journal*, v.71, n.1, pp. 43-55, 2003.

RECOLECTOR DE CIENCIA ABIERTA (RECOLECTA). (2016). recolecta.fecyt.es/

REES, H.L.; EGGLETON, J.D.; RACHOR, E.; VANDEN, Berghe, E. (2007) Structure and dynamics of the North Sea Benthos. *ICES Cooperative Research Report* 288. Copenhagen. 259 pp.

REES, T.; Zhang, Y. (2007) Evolving concepts in the architecture and functionality of OBIS, the Ocean Biogeographic Information System. In: *Proceedings of Ocean Biodiversity Informatics: An International Conference on Marine Biodiversity Data Management Hamburg, Germany, 29 November – 1 December, 2004* (eds. E. Vanden Berghe, et al.), pp. 167–176. COI Workshop Report, 202, VLIZ Special Publication 37.

REFRACTIONS Research. (2009). uDig User-friendly Desktop Internet GIS. http://udig.refractions.net/

REGISTRY of Research Data Repositories. (2016). http://www.re3data.org

REILLY, Susan, et. al. (2011). Report on integration of data and publications. Opportunities for data exchange. http://migre.me/qJaVT

REITZ, Joan M. Dictionary for Library and Information Science. (2016). *Libraries Unlimited.* http://www.abc-clio.com/ODLIS/odlis_about.aspx

RESEARCH Information Network. Research Funders' Policies for the Management of Information Outputs. *A report commissioned by the Research Information Network*. (2007). http://rinarchive.jisc-collections.ac.uk/our-work/research-funding-policy-and-guidance/research-funders-policies-management-information-outpu\

RIDING the wave - How Europe can gain from the rising tide of scientific data – Final report of the High Level Expert Group on Scientific Data. (2010). ec.europa.eu/information_society/newsroom/cf/itemlongdetail.cfm?item_id=6204

ROCHE, Dominique; G. LANFEAR, Robert; BINNING, Sandra. A.; HAFF, Tonya. M.; SCHWANZ, CAIN, Kristal E.; KOKKO, Hanna.; JENNIONS, Michael. D. Jennions; KRUUK, Loeske. E. B. Troubleshooting Public Data Archiving: Suggestions to Increase Participation. *PLOS Biology*, vol. 12, n.1. jan. 2014.

RODRÍGUES, Eva M. *Reflexiones em torno a los metadatos: um nuevo reto para la organización de bibliotecas digitales*. In: Videoconferência promovida pela Universidade de São Paulo e Universidad Carlos III, 10 dic. 2002.

ROSETTO, Márcia. (2003). *Metadados e formatos de metadados em sistemas de informação: caracterização e definição*. São Paulo. 112 p. (Dissertação de mestrado apresentada ao Curso de Pós – Graduação da Escola e Comunicações e Artes da Universidade de São Paulo).

ROYAL SOCIETY. *Science as an open interprise*. The Real Society Science Policy Centre Report. n. 2. jun. 2012. Disponible en*:* < http://royalsociety.org/uploadedFiles/Royal_Society_Content/policy/projects/sape/2012-06-20-SAOE.pdf >. Acceso en: 31 ene. 2014.

RUAL, P. (1989). For a better XBT bathy-message onboard quality control, plus a new data reduction method. In *Proceedings of the Western Pacific International Meeting and Workshop on TOGA Coare*, Bondy, France: pp. 823–833

SANTOS, E.; MIRAGLIA, S. Arquivos abertos e instrumentos de gestão da qualidade como recursos para a disseminação da informação científica em segurança e saúde no trabalho. *Ci. Inf.*, Brasília, v. 38, n. 3, dez. 2009.

SATRA-LE BRIS, Catherine., Quimbert, Erwann., Treguer, Mickael; Louarit, Abdelaziz. (2013). Sextant: French spatial data infrastructure for marine environments. *IMDIS, International Conference on Marine Data and Information Systems*, Lucca, Italia.

SC-ADMb. About SC-ADM. (2016). http://scadm.scar.org/about.html

SC-ADMa. Data Centers. http://scadm.scar.org/data_centres

SC-ADMc. (2016). http://scadm.scar.org

SCAR. (2016). http://www.scar.org

SCARb. Standing Scientific Group on Life Sciences. (2016). http://www.scar.org/researchgroups/lifescience/

SCHAAP, Dick M.A.; Glaves, Helen. ODIP - Ocean data interoperability platform - developing interoperability pilot project 1. *Geophysical Research Abstracts*, v. 16, abr. 2014.

SCHUTTER, E. Data Publishing and Scientific Journals: The Future of the Scientific Paper in a World of Shared Data. *Neuroinformatics.* n. 8. pp. 151–153. 2010.

SCIENTIFIC DATA. (2016). http://nature.com/scientificdata

SCOR & IODE (2008) SCOR/IODE Workshop on Data Publishing, Oostende, Belgium, 17–19 June 2008. COI Workshop Report No. 207. Paris: UNESCO. 23 pp. Disponíble en: < http://www.scor-int.org/Publications/wr207.pdf >. Acceso en 13 dic. 2014.

SHI J; HOU J; JIAO Y. (2011). Observing thermohaline structure of Polar Ocean with XCTD launched from helicopter. In *Proceedings of the Twenty-first International Offshore and Polar Engineering Conference*, Maui, Hawaii: pp. 998–1002.

SILVA, Fabiano Couto Corrêa; DOMINGUES, Marcelo Vinicius de La Rocha, ZIMMER, Marilene; CABRAL, João Carlos Centurion Rodrigues. Produção científica Antártica: análise bibliométrica dos repositórios institucionais. In: XII Encontro Nacional de Pesquisa em Ciência da Informação, 2011, Brasília. *Anais*, 2011.

SILVA, N. R. *Visualização 3d de dados oceanográficos simulados*. (2006). Mestrado, Programa De Pós-Graduação Em Ciência De Computação, Pontifícia Universidade Católica Do Rio Grande Do Sul, Puc-RS.

SIMSION, G. (2007). Data Modeling Theory and Practice. Bradley Beach, NJ, Technics Publications.

Sistema de Información Ambiental para el Programa Biota/FAPESP (SinBiota). (2016). http://sinbiota.biota.org.br

SMIT, Gene. (2008). Tagging: people-powered metadata for the social web. Berkeley: New riders.

SNOWDEN, D., Belbeoch, M., Burnett, B., Carval, T., Graybeal, J., Habermann, T., Snaith, H., Viola, H. & Woodruff, S. (2010). Metadata management in global distributed ocean observation networks. *Proc. OceanObs'09: Sustained Ocean Observations and Information for Society (Vol. 2)*, Venice, Italy.

SOMERFIELD, P.J.; ARVANITIDIS, C.; VANDEN Berghe, E., *et al*. (2009) MarBEF, databases and the legacy of John Gray. *Marine Ecology Progress Series* 382, pp. 221–224.

SORRIBAS Cervantes, Jordi. (2012). La gestión de datos marinos desde la perspectiva de un centro de datos de investigación. *Advances in Research Data Management* (Barcelona, 10 mayo). GrandIR/Universitat Politècnica de Catalunya. http://www.grandir.com/en/tecnical-session/advances-in-research-data-management-in-spain

SORRIBAS Cervantes, Jordi; LADONA, Emili García; CHIC, Òscar; OLIVÈ, Joan. Consejo Superior de Investigaciones Científicas. Entrevistas en noviembre 2014.

SOUZA, J. (2008). Brasil na Antártida 25 anos de História. São Carlos: Vento Verde.

STARR, Joan; GASTL, Angela. isCitedBy: A Metadata Scheme for DataCite. *D-Lib Magazine*, Vol. 17, n. 1/2, Jan/Feb. 2011.

STOKSTAD, Erik. Proposed rule would limit fish catch but faces data gaps. *Science, v.* 320, p. 1706–1707. jun. 2008.

TANI, Alice; CANDELA, Leonardo; CASTELLI, Donatella. Dealing with metadata quality: The legacy of digital library efforts. *Information Processing & Management, v. 49*, p.1194-1205. 2013.

TAPIA, Carlos Techeira; ABARCA, Luis Ariz; GONZALES, Mauricio, et al. (2006). Instituto de Fomento Pesquero). Proyectos BIP. Abril. Disponible en: < http://goo.gl/S97uNG >. Acceso en: 01 mar. 2016.

TENOPIR, Carol; ALLARD, Suzie; DOUGLASS, Kimberly; AYDINOGLU, Arsev Umur; WU, Lei. Data Sharing by Scientists: Practices and Perceptions. *Plos One*. June, 2011.

THANOS, C. A vision for global research data infrastructures. *Data Science Journal*, Sept. 2013. pp. 71–90.

TORRES-SALINAS, Daniel; ROBINSON-GARCÍA, Nicolás; CABEZAS-CLAVIJO, Álvaro. Compartir los datos de investigación: introducción al Data Sharing. *El profesional de la información*, 2012, v. 21, n.2, pp. 173-184.

UHLIR, P. F., Chen, R.S., Gabrynowicz, J.I., & Janssen, K. (2009) Toward Implementation of the Global Earth Observation System of Systems Data Sharing Principles. *Journal of Space Law 35*, pp. 201-290.

UNIVERSITY of Cambridge (2010). UK Funding Councils: Data Retention and Access Policies. http://www.lib.cam.ac.uk/dataman/resources/ Incremental_Cambridge_ factsheet_UKfunders_data policies.pdf

UK Data Archive. Research Data Lifecycle. (2016). http://data-archive.ac.uk/ create-manage/life-cycle

US Geological Survey (USGS). (2016). Disponible en: < https://www.usgs.gov/ >. Acceso en: 07 sep. 2016.

VAN DEN EYNDEN, V.; L. Corti, M. Woollard, L. BISHOP, L.; L. Horton. 2011. Managing and sharing data: Best practice for researchers. Wivenhoe Park, Colchester, Essex, UK: UK Data Archive, University of Essex. Disponíble en: < http://data-archive.ac.uk/media/2894/managingsharing.pdf >. Acceso en: 29 nov. 2014.

VAN DER GRAAF, Maurits; WAAIJERS, Leo. (2011). A Surfboard for Riding the Wave. Towards a four country action programme on research data. *A Knowledge Exchange Report*. http://migre.me/qJaV8

VANDEN BERGHE, E., Appeltans, W., Costello, M.J. & Pissierssens, P. (eds.) (2007a) *Proceedings of Ocean Biodiversity Informatics:* An International Conference on Marine Biodiversity Data Management Hamburg, Germany, 29 November – 1 December, 2004. Paris, UNESCO/COI, VLIZ, BSH, 2007. vi + 192 pp.

VANDEN BERGHE, E., Claus, C., Appeltans, W., *et al*. (2009) MacroBen integrated database on benthic invertebrates of European continental shelves: a tool for large-scale analysis across Europe. *Marine Ecology Progress Series* 382, pp. 225–238.

VANDEN BERGHE, E., Rees, H.L. & Eggleton, J.D. (2007b) NSBP 2000 data management. In: *Structure and dynamics of the North Sea Benthos* (eds. H.L. Rees, J.D. Eggleton, E. Rachor, & E. Vanden Berghe), pp 7–20. Copenhagen: ICES Cooperative Research Report 288.

VILLA, R. Segurança internacional: novos atores e ampliação da agenda. Lua Nova, São Paulo, n. 34, dez. 1994.

WALTERS, Tyler; SKINNER, Katherine. (2011). *New Roles for New Times*. Association of Research Libraries.

WELLCOME TRUST (2010). *Policy on Data Management and Sharing*. http:// www.wellcome.ac.uk/About-us/Policy/Policy-and-position-statements/ WTD002753.htm

WHITE House Office of Management and Budget. (2013). Uniform Administrative Requirements, Cost Principles, and Audit Requirements for Federal Awards. https://www.federalregister.gov/articles/ 2013/12/26/2013-30465/uniform-administrative-requirements-cost-principles-and-audit-requirements-for-federal-awards

WMO (2009). Report of the CCI Expert Team on WCP Requirements for Metadata. Geneva, Switzerland, World Meteorological Organization

WOODLEY, Mary S. DCMI Glossary. (2016). http://dublincore.org/documents/ usageguide/glossary.shtml

WRIGHT, D. J.; BLONGEWICZ, M. J.; HALPIN, P. N. & Breman, J. (2007). Arc Marine: GIS for a blue planet. California: ESRI Press.

XU, C.; XINYAN Z.; DAOSHENG, D. (2008). *Ontology based semantic metadata for imagery and gridded data*. In: Proceedings of the ISPRS Congress, Beijing, China: pp. 743–748.

ZELLER, D., Froese, R. & Pauly, D. (2005) On losing and recovering fisheries and marine science data. *Marine Policy* 29, pp. 69–73.

ZIKOPOULOS, P.C., Eaton, C., deRoos, D., Deutsch, T., G. (2012). *Lapis. Understanding Big Data Analytics for Enterprise Class Hadoop and Streaming Data*. New York: McGraw-Hill.

ZORRILLA, R; et. al. Conceptual view representation of the brazilian information system on antarctic environmental research. *Data Science Journal*, Vol.13, n. 30, Oct. 2014.

3TU. Datacentrum, (2015). 3TU. Datacentrum. http://datacentrum.3tu.nl/. en/home

APENDICE A: Estudio de usuarios (cuestionario enviado para los investigadores)

Prezado pesquisador,

Estou realizando uma investigação sobre a gestão de dados oceanográficos e polares e parte dela está focada no panorama atual da realidade brasileira.
Reconhecendo a importância do seu conhecimento sobre o assunto, peço-lhe que responda as questões que seguem. O tempo estimado para finalizar todas as respostas é de 3 minutos; ainda assim é possível responder de acordo com sua disponibilidade, por etapas.
Estes resultados serão utilizados no Programa de Doutorado em Informação e Comunicação na Sociedade do Conhecimento da Universidade de Barcelona com o objetivo de identificar a forma como dados científicos estão sendo armazenados pela comunidade oceanográfica, bem como as demandas apresentadas para a gestão eficiente dos dados que produzem.
As informações obtidas serão utilizadas de forma anônima e apenas para fins acadêmicos, servindo como subsídio para o desenvolvimento de um repositório de dados que atenda a comunidade oceanográfica brasileira.
Este formulário ficará disponível para preenchimento impreterivelmente até o dia 20 de junho de 2015 e caso deseje reeditar alguma resposta será possível acessar o formulário novamente até a data limite.

Instruções básicas para preenchimento:
* As respostas podem ser de múltipla escolha e referente tanto unicamente as investigações particulares como também um laboratório ou centro de investigação (quando o respondente for representante do mesmo).
* No caso de nenhuma resposta se aplicar a maneira como armazena os dados das suas investigações, basta escrever "nenhum" na última alternativa de cada questão (no campo "outro").

Desde já agradecemos sua colaboração.
Pesquisador: Fabiano Couto Corrêa
Currículo Lattes: http://lattes.cnpq.br/4635807083312321
Telefone para contato: +34 622 500 234 e-mail: fabianocc@gmail.com
Diretor de Pesquisa: Prof. Dr. Ernest Abadal Falgueras (Universidade de Barcelona)

*Obrigatório

1 - PRODUÇÃO DOS DADOS DE PESQUISA *

Em qual área gera dados primários?

- Oceanografía Física
- Oceanografía Biológica
- Oceanografía Química
- Oceanografía Geológica
- Outro: _____

2 - CARACTERÍSTICAS DOS DADOS DE PESQUISA *
Os dados produzidos se encontram em formato digital?

( ) Sim
( ) Não

3 - TIPOS DE DADOS *
Que tipos de dados são produzidos?

( ) Observacionais
( ) Estatísticos
( ) Imagens (satélite, mapas, etc.)
( ) Fotografias
( ) Registros de saídas de campo
( ) Video
( ) Simulações
( ) Áudio
Outro: _____

4 - FORMATOS DOS DADOS *
Que formatos digitais utiliza para criar ou arquivar dados primários (brutos)?

( ) Planilha de cálculo (Ex. Excell)
( ) Portable Document Format (.pdf)
( ) Formatos de imagen (.bmp; .jpg; .tiff, etc.)
( ) Texto (.txt o .doc)
( ) Software estatística (Ex. SPSS)
( ) Formatos de áudio (MP3, wav, etc.)
( ) Arquivos de bases de datos (Ej. Access, MySQL)
( ) Rich Text Files (.rtf)
( ) Formatos de imagem vetorial (.cdr; .ai; etc.)
( ) Hypertext markup language (HTML)
( ) Sistema de Informação Geográfica (SIG)
( ) Extensible markup language (XML)
Outro: _____

5 - Quais formatos de dados oceanográficos utilizas para criar ou arquivar dados primários? *

( ) XML
( ) KML
( ) CSV
( ) GRIB
( ) Audio
( ) Matlab
( ) GeoTIFF
( ) JSON
( ) NetCDF
( ) ASCII
( ) Shapefiles
( ) HDF
Outro: _____

6 - En relação as normas de metadados para informação oceanográfica, quais utiliza? *

( ) ISO 19115 - Informação geográfica
( ) Marine Community Profile
( ) Common Data Index
( ) Directorio Interchange Form
( ) Geonetwork
( ) Relatórios resumidos de cruzeiros
( ) Cruise Summary Report (CSR)
( ) Base de dados CSR/ROSCOP
( ) Amostra CSR Record
Outro: _____

7 - APLICAÇÕES DE SOFTWARE *
Quais aplicações de software utiliza para gerenciar os dados das suas investigações?

( ) Software para transcrição de áudio e vídeo
( ) Software para edição de imagens
( ) Software para tratamento de dados qualitativos
( ) Software para tratamento de dados quantitativos
Outro: _____

8 - ALTERNATIVAS DE COMPARTILHAMENTO DOS DADOS *
De qual maneira demais pesquisadores podem obter acesso aos dados das suas investigações?

( ) Em formato impresso
( ) Dados distribuídos por email
( ) Dispositivos de armazenamento portáteis (Ex: CD, Pen Drive, etc.)
( ) Realizo download para um servidor web e disponibilizo acesso mediante solicitação
( ) Realizo download para um servidor web possibilitando acesso público (Ex: repositórios institucionais)
( ) Deixo todos os dados relacionados a um artigo publicado sob responsabilidade e decisão do editor de uma revista
( ) Não é possível acessar os dados das minhas investigações
Outro: _____

9- Em relação aos repositórios que deposita os dados levantados em suas pesquisas, assinale qual(ais) são utilizado(s)
Caso nenhuma reposta se aplica a sua situação, passe a seguinte questão sem assinalar nenhuma alternativa ou então indique o nome da base de dados em "Comentários adicionais”

REPOSITORIO	Consultar	Upload dos dados	Download dos dados
SeaDataNet			
Systèmes d'Informations Scientifiques pour la MER (SISMER)			
British Oceanographic Data Centre (BODC)			
Geological and Geophysical Data (Geo-Seas)			
European Marine Observation and Data Network (EMODnet)			
JERICO			
Rolling Deck to Repository			
National Oceanic and Atmospheric Administration (NOAA)			
Australian Ocean Data Center Facility (AODC)			
Integrated Marine Observing System (IMOS)			
Intergovernmental Oceanographic Commission			

10 - MOTIVAÇÕES PARA COMPARTILHAR SEUS DADOS *
Quais fatores são considerados motivacionais para o compartilhamento dos dados científicos das suas investigações em um repositório digital?

( ) Requisito por parte da agência que investiga minha pesquisa
( ) Benefícios para minha carreira profissional (méritos para minha avaliação institucional, etc)
( ) Benefícios para instituição que trabalho
( ) Benefícios para comunidade científica
( ) Benefícios potenciais para sociedade em geral
( ) Permitir colaborações e contribuições de demais investigadores

( ) Expor os resultados obtidos para validação pelos meus pares
( ) Ampliar a visibilidade das minhas pesquisas
Outro:_____

11 - FATORES DESANIMADORES *

Quais fatores podem ser considerados desanimadores para compartilhamento dos dados levantados nas suas pesquisas?
( ) Falta de financiamento da área que desenvolvo investigações
( ) Tempo e esforço necessário para compartilhamento
( ) Desconhecimento das normas e procedimentos para compartilhar meus dados
( ) Restrições éticas
( ) Restrições relativas a segurança e confidencialidade dos dados
Outro:_____

12 - SERVIÇOS DE APOIO *
Você está interessado em contar com serviço de ajuda para compartilhamento dos seus dados científicos? Selecione todas as alternativas que se apliquem:
( ) Assessoramento sobre questões práticas relacionadas com gestão de dados
( ) Assessoramento sobre opções para armazenar, gestionar e compartilhar de forma segura
( ) Assessoramento sobre preservação dos dados levantados em minhas investigações
( ) Assessoramento sobre a criação de um plano de gestão de dados científicos
Outro:_____

13 - se desejar participar de etapa futura desta pesquisa, por favor deixe seu nome, instituição e email:

14 - Espaço para comentários adicionais:

15 - Identifique o nome da instituição que trabalha: *

APENDICE B: The ArcGis Marine Data Model

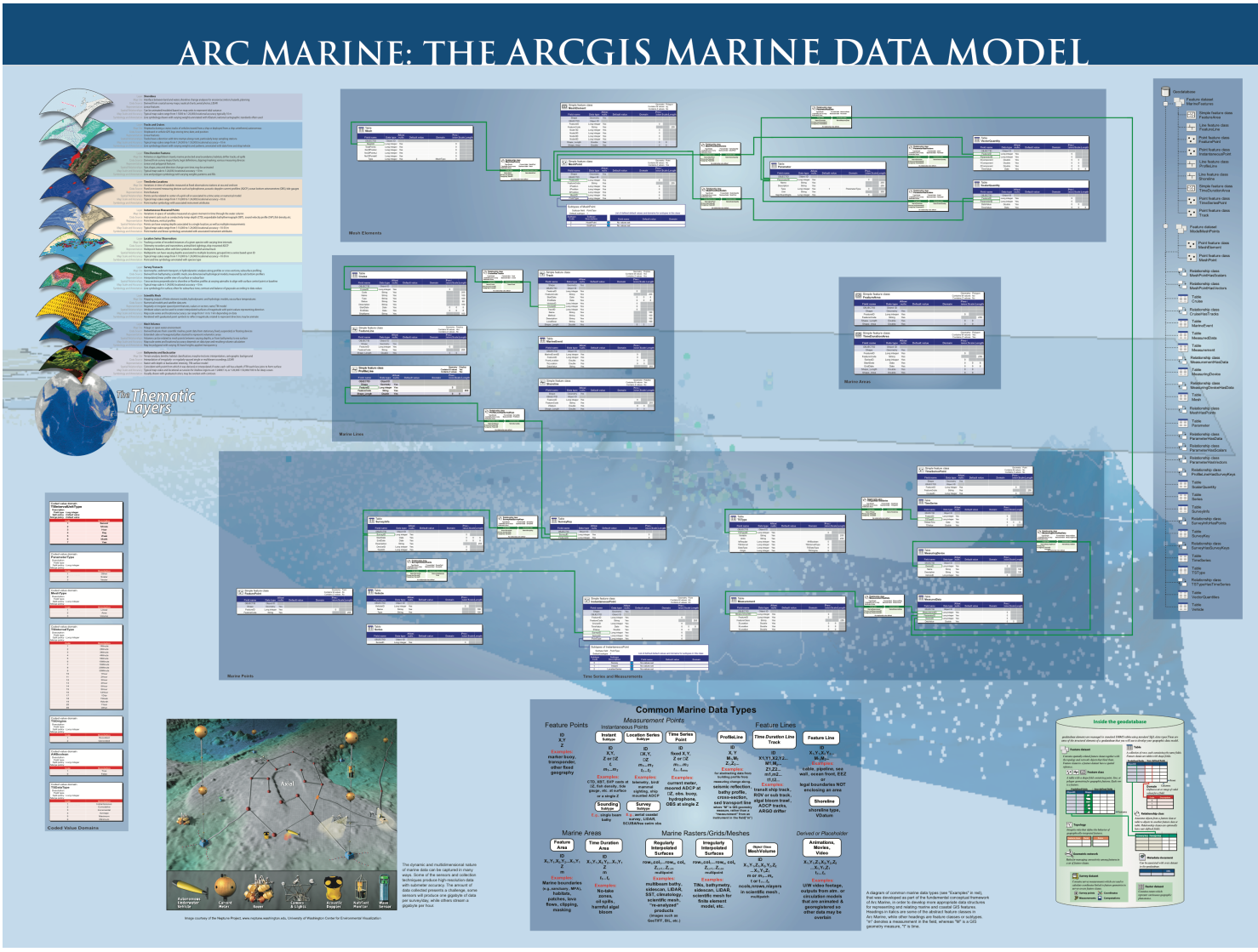

Figura 35: The ArcGis Marine Data Model