PANGAEA® als vernetztes Verlags- und Bibliothekssystem für wissenschaftliche Daten

Dr. Michael Diepenbroek, MARUM, Universität Bremen
Dr. Hannes Grobe, Alfred-Wegener-Institut für Polar- und Meeresforschung

Abstract

Since 1992 PANGAEA® serves as an archive for all types of geoscientific and environmental data. From the beginning the PANGAEA group started initiatives and aimed at an organisation structure which – beyond the technical structure and operation of the system – would help to improve the quality and general availability of scientific data. Project data management is done since 1996. 2001 the *World Data Center for Marine Environmental Sciences* (WDC-MARE) was founded and since 2003 – together with other German WDC – the group was working on the development of data publications as a new publication type. To achieve interoperability with other data centers and portals the system was adapted to global information standards. PANGAEA® has implemented a number of community specific data portals. 2007 – under the coordination of the PANGAEA® group – an initiative for networking all WDC was started. On the long-term ISCU plans to develop the WDC system into a global network of publishers and libraries for scientific data.

Abstract

Seit 1992 werden mit PANGAEA® wissenschaftliche Daten aus allen Bereichen der Geowissenschaften und Biologie archiviert. Von Beginn an wurden Initiativen gestartet und eine Organisationsstruktur angestrebt, welche über das technische System hinaus helfen die Qualität und Verfügbarkeit von wissenschaftlichen Daten zu verbessern. Seit 1996 wird intensiv Projektdatenmanagement betrieben. 2001 wurde das *World Data Center for Marine Environmental Sciences* (WDC-MARE) gegründet und ab 2003 zusammen mit den weiteren deutschen WDC an der Entwicklung der Datenpublikation als neuem Publikationstyp gearbeitet. Um eine Interoperabilität mit anderen Datenzentren und Datenportalen zu erreichen wurde das System gleichzeitig an globale Informationsstandards angepasst. PANGAEA® hat in der Folge selbst eine Reihe von community spezifischen Datenportalen implementiert. Koordiniert von der PANGAEA® Gruppe wurde 2007 eine Initiative zur Vernetzung der WDC gestartet. Langfristig ist vorgesehen das WDC System in einen globalen Verlags- und Bibliiotheksverbund für wissenschaftliche Daten zu entwickeln.

Einleitung

Datenzentren, eine neuzeitliche Erfindung, entstanden zumeist mit der Motivation, wissenschaftliche Daten, die regelmäßig in einem gewissen institutionellen oder projektgebundenen Rahmen erhoben werden, langzeitlich zu sichern. So war das Geophysikalische Jahr 1957 Ausgangspunkt für die Gründung einer Reihe von global verteilten Weltdatenzentren, die sich der in diesem und in den nachfolgenden Jahren produzierten wissenschaftlichen Daten annehmen und dauerhaft archivieren sollten. Das System der *World Data Center* (WDC) ist an den *International Council for Scientific Unions* (ICSU) gebunden und feierte in diesem Jahr seinen 50. Geburtstag. Seit einigen Jahren unterliegt das System einem zunehmenden Erneuerungsdruck. Die exponentiell steigende Datenflut und die Entwicklung des Internet führte zum Aufbau vieler neuer Daten haltender Systeme, eines davon ist das 1992 entstandene *Publishing Network for Geoscientific and Environmental Data* (PANGAEA® – www.pangaea.de). 2001 gründete die PANGAEA® Gruppe das ICSU *World Data Center for Marine Environmental Sciences* (WDC-MARE – www.wdc-mare.org).

PANGAEA® war von Anfang an als breitspektral arbeitendes System gedacht. Die Heterogenität und Dynamik der Geowissenschaften und Biologie erforderte ein möglichst flexibles System zur Erfassung, Bearbeitung und Archivierung der vielfältigen Daten. Dennoch wurde bereits in der Aufbauphase klar, dass ein gutes technisches System zwar Grundvoraussetzung ist, nicht aber die prinzipiellen Probleme der Verfügbarkeit und Qualität wissenschaftlicher Daten lösen kann. Wissenschaftliche Primärdaten sind, neben den Publikationen, das zweite wichtige Ergebnis, das langfristig und in nachnutzbarer Form nach dem Prinzip des offenen Zugangs (DFG 1998, ESF 2000, Berliner Erklärung 2003, OECD 2004) verfügbar sein muss. In den 70er Jahren des vorigen Jahrhunderts war es noch üblich, Primärdaten direkt in einer Publikation oder im Anhang einer Zeitschrift abzudrucken. Mit steigenden Datenmengen und dem Übergang zur elektronischen Publikation wurde diese Praxis, primär aufgrund der Kosten, aufgegeben. Wissenschaftliche Verlage erlauben zwar die Ablage von zu einer Publikation gehörenden Primärdaten, die Archivierung folgt jedoch keinerlei Standards oder einheitlichen Strukturen und ist vom peer review ausgeschlossen, kann also nicht als Vorbild für eine allgemeine Lösung des Problems gelten.

Viele Datenzentren, auch ein guter Teil der ICSU WDC sind dem gegenüber technisch gut vorbereitet. Dennoch folgt auch hier die Archivierung zumeist keinen globalen Standards.

Die Trennung von wissenschaftlicher Publikation und zugrunde liegenden Primärdaten kann als gravierendes strukturelles Problem in den empirischen Wissenschaften gesehen werden. Nicht nur die Evaluierung einer Publikation auch die Nachnutzung der Ergebnisse ist erheblich eingeschränkt. Es gibt keine wirklich autorisierten und authentifizierten Orte für die langzeitliche Aufbewahrung von wissenschaftlichen Daten, keine Korrelation zwischen archivierten Daten und

wissenschaftlicher Publikation und keine Vernetzung zwischen den Datenzentren. Gebraucht wird ein globales Verlags- und Bibliothekssystem für die Erfassung, Archivierung und Publikation wissenschaftlicher Daten. Hier spielen die ISCU WDC eine aktive Rolle. Die drei deutschen WDC (WDC-Climate, WDC-RSAT, WDC-MARE) und das GFZ haben – zusammen mit der Technischen Informationsbibliothek (TIB), Hannover - in den letzten drei Jahren ein praktikables System zur Publikation wissenschaftlicher Daten geschaffen (Schindler et al 2005, www.std-doi.de). WDC-MARE mit dem Informationssystem PANGAEA® und seinem redaktionellen System kann dabei bereits jetzt als Referenz für ein Verlags- und Bibliothekssystem für wissenschaftliche Daten gesehen werden und soll im weiteren näher beschrieben werden.

Von der Datenerfassung zur Publikation

WDC-MARE / PANGAEA® wird als permanente Einrichtung vom Zentrum für Marine Umweltwissenschaften (MARUM) der Universität Bremen und der Stiftung Alfred-Wegener-Institut für Polar- und Meeresforschung (AWI) in Bremerhaven betrieben. 3 Wissenschaftler sind für die grundsätzliche Organisation und Entwicklung zuständig. Ein Team von durchschnittlich 6-8 Wissenschaftlern übernehmen seit 1996 das Projektdatenmanagement (www.pangaea.de/Projects). Die daraus gewonnenen Drittmittel tragen zu einem erheblichen Teil zur Finanzierung des Betriebes bei.

Datenerfassung, Qualitätssicherung, redaktionelle Bearbeitung und Archivierung

Die Akquise wissenschaftlicher Daten ist ein gravierendes Problem. Eigenen Schätzungen zufolge werden nur wenige Prozent der global produzierten wissenschaftlichen Daten auch in geeigneten Datenbanken langfristig archiviert. Es kommt selten vor, dass Daten freiwillig Datenzentren überlassen werden. Im institutionellen Rahmen gibt es seit einigen Jahren eine Aufbewahrungspflicht (DFG 1998). Ebenso sind viele wissenschaftliche Projekte bzw. Programme mit entsprechenden Auflagen versehen. Absprachen in solchen Kontexten erleichtern die Datenerfassung, können das Problem jedoch nicht vollständig beseitigen.

Als relativ effektiv hat sich Datenmanagement als finanzierter Bestandteil wissenschaftlicher Projekte erwiesen (http://www.pangaea.de/projects/). Diese Art der Projektförderung ist zunehmend in Verbundprojekten von BMBF und DFG zu beobachten; eine allgemeine Regelung steht jedoch noch aus. Auf der EU Ebene ist Datenmanagement seit seit den ersten Förderprogrammen wichtiges Kriterium für die Bewertung von Projekten (MAST Data Management Code). Projekte wie z.B. CARBOOCEAN (http://www.carboocean.org/), die eine verbesserte Quantifizierung von Kohlenstoffbilanzen im marinen Bereich zum Ziel haben, ist die möglichst vollständige Erfassung qualitätsgesicherter Daten notwendige Bedingung für den Erfolg des Projekts. Allgemein bauen komplexe und/oder großskalige Ansätze in der

Global Change Forschung auf die Daten und Ergebnisse vieler kleiner Projekte und sind oft nur so realisierbar.

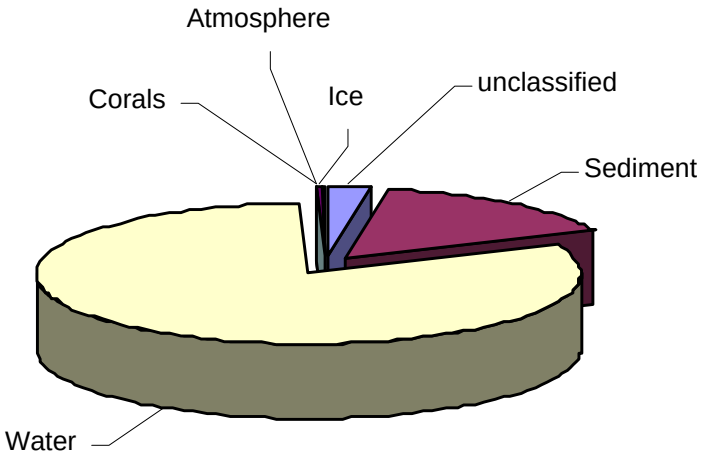
Die PANGAEA® Gruppe betreibt Projektdatenmanagement seit mehr als 10 Jahren. Dies ist die wichtigste Quelle für neu zu archivierende Daten, wesentlich bedingt durch die Nähe mit den Wissenschaftlern. Zudem trägt Projektdatenmanagement auch erheblich zur Finanzierung des Betriebes von PANGAEA® bei. Dies schafft Kapazitäten, die es der Gruppe ermöglichen auch nicht finanzierte Projekte wie z.B. die globale abschließende Erfassung und Publikation von Daten aus dem IGBP Project Joint Global Ocean Flux Studies (JGOFS) durchzuführen (Sieger et al 2005).

Unverzichtbarer Bestandteil des Datenmanagements ist die Qualitätssicherung. Kernaussage ist dabei, dass es weniger auf die Qualität der Daten ankommt als vielmehr, dass diese durch den Nachnutzer der Daten einschätzbar wird. Wichtig ist also die vollständige Erfassung von beschreibenden Informationen, den Metadaten, und die Einhaltung entsprechender Standards (s. nächster Abschnitt). Minimal zu beantwortende Fragen sind: Wer hat was wo wann und wie gemessen oder beobachtet. Darüber hinaus wird bei PANGAEA® standardmäßig die Validität der eingesetzten Methoden geprüft und ob z.B. die Präzision der Daten mit der eingesetzten Methodik korrespondiert. Ausreißer werden ermittelt und entsprechend markiert. Für die eigentliche Qualität der Daten steht der Datenproduzent mit seinem Namen und/oder seiner Institution.

Die Bearbeitung und Archivierung von Daten variiert im Allgemeinen mit den Datenzentren und Datentypen. Es gibt praktisch weder einheitliche Datenmodelle noch allgemein nutzbare redaktionelle Systeme. Üblich ist die Nutzung relationaler Datenbanken im Serverbereich, was insbesondere für die Metadaten eine gewisse Konsistenz gewährleistet.

Daten für ca. 30000 Parameter, z.B.:

Sediment- und Eis Profile
Seismische Profile
Armosphärenprofile
Ozeanprofile
Mineralverteilungen
Geologische Karten
Bilder und Filme
Zeitreihen
Plankton und Fisch



Datensätze: 551 012
Datenobjekte: 1 823 824 468

Abbildung 1 Inhalte von PANGAEA®. Stand 9/2007

Das WDC-MARE mit dem Informationssystem PANGAEA® archiviert seit 1992 georeferenzierte Daten primär aus der Meeres- und Umweltforschung (Diepenbroek et al 2002). Derzeit sind ca. 550.000 Datensätze mit knapp zwei Milliarden Mess- und Observationsdatenpunkten verfügbar. Sie sind verknüpft mit ca. 30.000 unterschiedlichen Datentypen (Parameter, Variable), ca. 6000 Autoren (Datenurheber), knapp 6000 klassischen Publikationen und mehr als 300.000 Probenlokationen. Der jährliche Zuwachs beträgt durchschnittlich mehr als 20% des Gesamtbestands.

In PANGAEA® werden Daten und Metadaten systematisch über ein redaktionelles System erfasst und bearbeitet. Das System steigert erheblich die Effizienz bei der Pflege der Daten und hilft, Fehler zu vermeiden. Bei kleinen Datenzentren mit relativ spezialisierten Dateninhalten ist ein redaktionelles System meist noch entbehrlich. PANGAEA® war jedoch von Beginn an als System gedacht um alle Arten von geowissenschaftlichen und biologischen Daten zu archivieren.

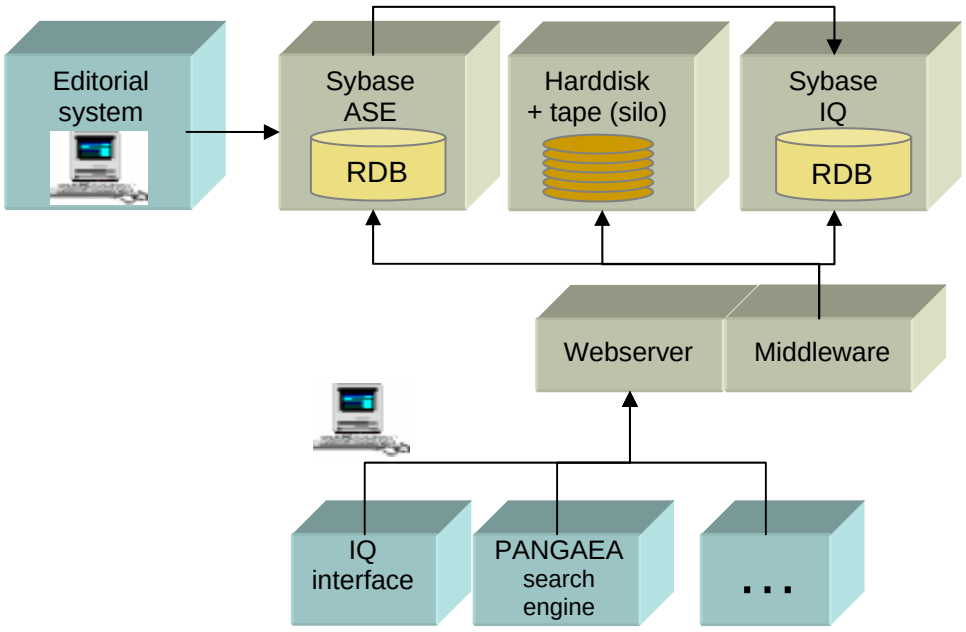


Abbildung 2 Technischer Aufbau von PANGAEA®

Für die Archivierung von Daten und Metadaten wird ein Relationales Datenbank Managementsystem (RDBMS - Sybase) genutzt. Backups werden an physikalisch unterschiedlichen und gesicherten Orten abgelegt. Damit ist der Datenbestand weitestgehend vor Vorlust geschützt. Abbildung 2 zeigt vereinfacht den technischen Aufbau von PANGAEA®. Massendaten, wie z.B. Geophysik, und binäre Objekte, wie z.B. Bilder werden auf einem Festplattenarray abgelegt und migrieren ggf. in ein Bandsilo. Sybase IQ ist ein Datenwarenhaus, in dem sämtliche numerischen und textuellen Datenwerte gespiegelt werden. Das Datenwarenhaus ermöglicht

performante retrievals von beliebigen raumzeitlichen Ausschnitten. Die Metadaten werden als Kompilat an das Suchergebnis angehängt.

Datenpublikation

Zusammen mit weiteren WDC in Deutschland hat die PANGAEA® Gruppe innerhalb des DFG-Projekts „Zitierfähigkeit und Publikation wissenschaftlicher Primärdaten“ in den letzten 3 Jahren das Konzept zur Publikation wissenschaftlicher Daten entwickelt und prototypisch implementiert (Schindler et al 2005, Klump et al 2006). Gemeinsam wurden allgemeine Anforderungen an diesen neuen Publikationstyp untersucht:

- Der formale Aufbau der Publikation, d.h. welche beschreibenden Elemente sind notwendig und welche optional, wie müssen diese gestaltet werden, welche Datenformate und Standards sind sinnvoll?
- Die Granularität von publizierbaren Datensätzen
- Die Notwendigkeit „peer review“ ähnlicher Verfahren zur Qualitätssicherung
- Die Anforderungen an die Datenarchive bezüglich Langzeitverfügbarkeit und dauerhafter eindeutiger Referenzierbarkeit von publizierten Datensätzen mittels persistenten Identifizierern wie dem ‚Digital Object Identifier’ (DOI). Für die Zertifizierung von Datenzentren wurde neben dem eigenen Erfahrungshintergrund auch wesentlich auf das OAIS Referenzmodell (NASA 2002) und die Ergebnisse des BMBF Projekts NESTOR (http://www.langzeitarchivierung.de/) zurückgegriffen.

Die Ergebnisse wurden in den beteiligten Einrichtungen als Richtschnur genutzt, um einerseits die Strukturen der Datenarchive entsprechend anzupassen, andererseits für neue Datensätze und die Pflege der vorhandenen Datensätze redaktionelle Organisationsschemata zu entwickeln. Derartige Schemata wurden in allen Einrichtungen prototypisch verwirklicht, sind jedoch – nach Einrichtung verschieden – erst mehr oder weniger vollständig in die technische Umgebung und den wissenschaftlichen Arbeitsprozess integriert worden.

Hier ist insbesondere das oben beschriebene Problem der Granularität zu nennen. In der bisherigen Arbeit haben sich die beteiligten Einrichtungen auf ein einfaches Modell geeinigt, indem individuelle publizierbare Datenentitäten im wissenschaftlichen Archiv zu zitierfähigen Datenentitäten zusammengefasst werden. Zitierfähige Datenentitäten repräsentieren die Schnittstelle zwischen Datenarchiv und wissenschaftlicher Literatur. Einerseits können Datensätze als publizierbare Entität direkt zitierfähig sein (wie Beispiele aus der Erdsystemforschung und aus dem medizinischen Bereich zeigen), andererseits kann es sein, dass man Datensätze – abhängig von der Nutzung – auf unterschiedlichen Ebenen zitieren möchte. So treten z.B. Fälle auf, bei denen man auf eine komplette Datensammlung verweisen möchte und andere Fälle, bei denen man auch einzelne Datensätze aus einer Sammlung zitieren möchte. Hier wird es notwendig sein, unterschiedliche, von den Datentypen

abhängige Modelle zu entwickeln, um eine flexiblere Nutzung und Zitierweise zu ermöglichen.

Als weiteres Problem kommt hinzu, dass für die großen heterogenen Datenarchive (wie WDC-MARE, GFZ) ein beträchtlicher initialer Aufwand nötig ist, um die in mehr als 10 Jahren gewachsenen Bestände entsprechend zu bearbeiten und bereits registrierte Datensätze als zitierfähig auszuweisen oder zu zitierfähigen Einheiten zusammenzufassen. Jeder zu publizierende Datensatz erfordert Rücksprache mit den Datenurhebern oder weiteren Wissenschaftlern aus dem jeweiligen Forschungsumfeld und ggf. manuelle Korrekturen der Metadaten. In allen Einrichtungen wurden erste Ansätze eines „peer reviews“ erprobt.

Es zeigt sich, dass die verschiedenen Datenwelten einen sehr unterschiedlichen Aufwand beinhalten. Im WDC-C sind eine überschaubare Anzahl an relativ homogenen aber extrem hochvolumigen Simulationsdatensätzen archiviert, WDC-MARE und GFZ stellen dagegen eine große Anzahl an äußerst heterogenen Observations- und Messdaten zur Verfügung, die im GFZ zudem noch über mehrere Datenbanken verteilt sind.

Die bisher nur in Ansätzen bewältigte Aufarbeitung der „Altbestände“ im Sinne des neuen Review- und Publikationsprozesses und die noch wenig etablierten Arbeitsabläufe führten dazu, dass zwar nahezu alle Datensätze in der DOI registry an der TIB erfasst sind, bislang jedoch weniger als 1000 von den Autoren zertifizierte und damit zitierfähige Einträge in TIBORDER, dem Online-Katalog der TIB, sichtbar sind. Nichtsdestotrotz sind alle Datensätze qualitätsgeprüft, vollständig mit Metadaten annotiert, direkt über die DOI zugreifbar und unterliegen denselben Anforderungen an die Langzeitarchivierung wie die von den Autoren zertifizierten und zitierfähigen Einträge. Künftig wird es jedoch nicht mehr zulässig sein, Datensätze ohne eine eindeutige Zuordnung zu einer zitierfähigen Datenentität zu registrieren. Die entsprechende Aufarbeitung der bereits registrierten und neu hinzukommenden Datensätze wird als langfristige Eigenleistung weitergeführt.

Auf diesem Hintergrund waren zu Beginn des Datenpublikationsprojektes vor einigen Jahren die vordringlichste Aufgaben, (1) die Überarbeitung der Metadatenstrukturen und -bestände und (2) die technische Anpassung des vorhandenen Systems zur redaktionellen Bearbeitung von Daten und Metadaten.

Nach einer zügigen Anpassung der Metadatenstrukturen wurden eine Reihe von Routinen geschrieben, mit deren Hilfe ein großer Teil der für die Datenpublikation relevanten Metadaten harmonisiert wurden. So wurden damit z.B. weitgehend Doppeleinträge von Autoren beseitigt und die Schreibweise vereinheitlicht. Die Korrektur von fraglichen und fehlenden Einträgen, sowie z.B. auch die Vergabe und Korrektur von Titeln mussten händisch nachgeführt werden. Der Prozess wurde 2006 nahezu abgeschlossen.

Für den Import neuer Daten und die Metadatenbearbeitung nutzt PANGAEA® ein in mehr als 10 Jahren gewachsenes, global nutzbares System (client/server), welches

die manuellen Arbeitsschritte auf ein Minimum reduziert. Die Anpassung dieses Systems an ein redaktionelles System zur Datenpublikation ist ein laufender iterativer Prozess, an dem sowohl die Systemmanager als auch die mit PANGAEA® arbeitenden Datenkuratoren intensiv beteiligt sind. So war es neben einer Vielzahl von Anpassungen auch notwendig, einen zeitlichen Ablauf einzuarbeiten. Neu importierte Datensätze werden danach nicht sofort registriert, sondern sind in einem Zeitraum von 28 Tagen noch änderbar bzw. vollständig ersetzbar. Nach Ablauf dieser Zeit werden sie automatisch registriert, ggf. als zitierfähig ausgewiesen und sind dann bis auf einige beschreibende Elemente statisch. Die Registrierung wurde flexibel in die bestehende Infrastruktur integriert (Abb. 2).

Insgesamt ist die für ein Publikationssystem notwendige Umstellung weitgehend abgeschlossen, so dass der redaktionelle Arbeitsaufwand – abgesehen von der vermehrten Kommunikation - insgesamt etwa gleich geblieben ist; ein nicht unwesentlicher Aspekt bezogen auf die laufenden Betriebskosten. Unvollständig ist jedoch noch die Ausweisung von Datensätzen als zitierfähige Datensätze, primär wegen der überwiegend noch fehlenden Zertifizierung der Datenpublikationen durch die Autoren aber auch wegen des noch nicht ausgearbeiteten Reviewverfahrens und des mehrfach angesprochenen Granularitätsproblems. Mit Blick auf die Akzeptanz von Datenpublikationen und die vom WDC-MARE zu erwartende Menge zitierfähiger Datensätze erfordert das weitere Vorgehen jedoch Zeit und Sensibilität. Beispiele zitierfähiger Datensätze sind:

http://doi.pangaea.de/10.1594/PANGAEA.472287
http://doi.pangaea.de/10.1594/PANGAEA.472492
http://doi.pangaea.de/10.1594/PANGAEA.472492

Standards, Vernetzung und Portale

Die Vernetzung von Datenproduzenten, -archiven, und -konsumenten – allgemein als Geodateninfrastrukturen bezeichnet - ist eine notwendige Vorbedingung für komplexe oder großskalige Datenkompilationen. *Global Spatial Data Infrastructures* (GSDI - http://www.gsdi.org/) ist eine Vision, die in vielerlei Hinsicht erheblich zur Effizienz bei der Zusammenstellung und Auswertung von wissenschaftlichen Daten beitragen würde. Mit der Zustimmung von mehr als 60 Ländern zum 10-Jahres Implementierungsplans eines *Global Earth Observing System of Systems* (GEOSS) der *Group on Earth Observations* (GEO - http://www.earthobservations.org/) wurde 2005 erstmals auf der ministeriellen Ebene ein wirksamer Rahmen geschaffen, der die Vernetzung und Bereitstellung von wissenschaftlichen Daten fördern soll. Nichtsdestotrotz ist GEOSS aufgrund fehlender Ressourcen auf existierende Kapazitäten und Aktivitäten angewiesen. Während des diesjährigen Treffens der WDC Direktoren wurde daher eine Initiative zur Vernetzung der WDC beschlossen. Dies wird von den WDC nicht nur als sinnvoller Beitrag zu GEOSS gesehen sondern auch als erster Schritt zur Schaffung eines globalen Verbunds von

Datenbibliotheken. Die Vernetzungsinitiative wird von der PANGAEA® Gruppe koordiniert.

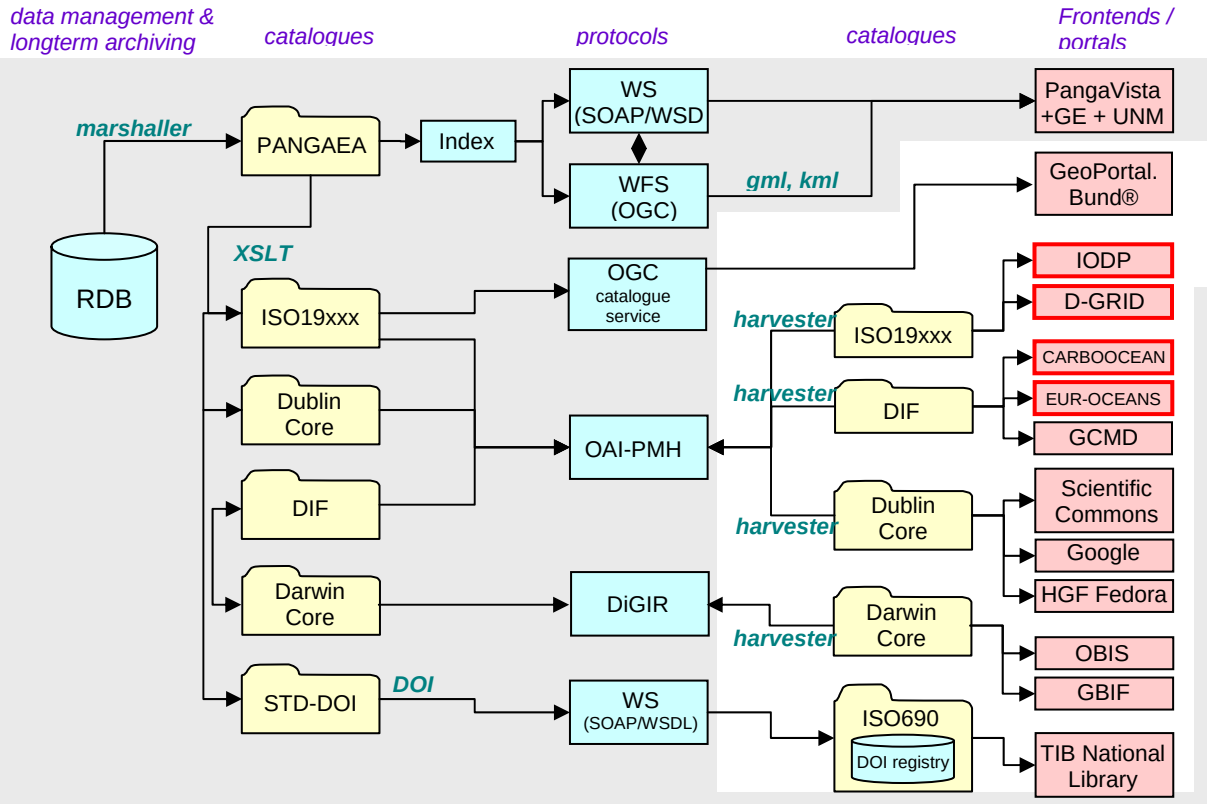


Abbildung 3 Metadaten-Infrastruktur von PANGAEA®. Grau hinterlegte Teile gehören vollständig zur internen Infrastruktur von PANGAEA®. Rot umrandet sind Portale, die weitestgehend von der PANGAEA® Gruppe implementiert wurden.

Die Gruppe hat in den letzten 5 Jahren systematisch an der Vernetzungsfähigkeit des Systems gearbeitet und stellt mittlerweile eine Vielzahl unterschiedlicher, allgemein nutzbarer Dienste zur Verfügung. Die Implementierung internationaler Informations- und Geodatenstandards spielt dabei eine wesentliche Rolle. Wichtige Instanzen sind hierbei ISO, OGC und W3C. Auf der Metadatenebene werden für jeden Datensatz die in der RDB gespeicherten Informationen über eine ‚marshalling' Routine zunächst in einem proprietärem Metadatenschema gesammelt und dann per XSLT in die diversen Inhaltsstandards übertragen. Dazu gehören z.B. ISO19115 als mittlerweile wichtigster Standard und das ‚Directory Interchange Format' (DIF - http://gcmd.nasa.gov/User/difguide/), Wichtige Protokolle sind der OGC Catalogue Service (http://www.opengeospatial.org/standards/cat) und das Open Archives Initiatives Protocol for Metadata harvesting (OAI-PMH - http://de.wikipedia.org/wiki/Open_Archives_Initiative). Letzteres ist relativ einfach zu implementieren und findet breite Verwendung nicht nur in der Bibliothekswelt, sondern zunehmend auch bei Datenzentren. Einen Überblick gibt Abbildung 3.

So hat die PANGAEA® Gruppe eine Reihe von community- und projektspezifischen Portalen implementiert. Das Portalframework ist generisch und basiert auf den Komponenten harvester, Indizierer mit Suchmaschine (Lucene - http://lucene.apache.org/java/docs/) und API (Schindler et al 2007, http://www.panfmp.org/). Beispiele sind das Portal für das *International Ocean Drilling Program* (IODP - http://sedis.wdc-mare.org/) und für die EU Projekte EUR-OCEANS (http://www.eur-oceans.eu/integration/wp2.2/) und CARBOOCEAN (http://dataportal.carboocean.org/)

Da informationstechnische Entwicklungen einer erheblichen Dynamik unterliegen, hat sich die PANGAEA® Gruppe bewusst darauf eingestellt, eine Vielzahl von Standards bedienen zu können.

Z.Z. beschränkt sich der standardisierte Informationsaustausch jedoch noch fast ausschließlich auf Metadaten. Dies liegt vor allem an dem Fehlen praktikabler und allgemein akzeptierter Datenaustauschformate. Lediglich für das Deutsche Community GRID C3 (http://www.c3grid.de/) werden von PANGAEA® auch Daten zur Verfügung gestellt.

Im Bereich des Datenaustausches sind GRID-Architekturen üblich, die derzeit jedoch noch sehr auf spezielle Datentypen und Dienste fokussiert sind. Für allgemeine und leicht zu implementierende Architekturen besteht noch Entwicklungsbedarf. So ist ein spezielles Problem die Verwendung einheitlicher Begriffe in den Metadaten zur Steuerung von Applikationen. Hier fehlen ebenfalls internationale Standards. Über die Europäischen Initiative INSPIRE (http://de.wikipedia.org/wiki/INSPIRE) besteht die Hoffnung, dass solche Vokabularien aufgebaut werden. Dies ist jedoch als langfristige Aufgabe zu sehen.

Fazit

Mit seinen langfristig angelegten und gesicherten Archivstrukturen, dem hocheffizienten redaktionellen System und der weitgehenden Interoperabilität mit Datenportalen und anderen Datenzentren hat sich PANGAEA® als exemplarisches Verlags- und Bibliothekssystem für wissenschaftliche Daten entwickelt.

Der im deutschen WDC Konsortium entwickelte und innerhalb von PANGAEA® realisierte Ansatz zur Publikation wissenschaftlicher Daten geht weit über die z.B. von der Human Genom Community praktizierte Kopplung von Datenarchivierung und wissenschaftlichem Aufsatz hinaus. Es ermöglicht Wissenschaftlern Datensätze eigenständig zu publizieren. Jede Datenpublikation wird mit einem aussagekräftigen Zitat und einem persistenten Identifizierer (DOI) versehen und erlaubt somit eine verlässliche Referenzierung. Die Zitierfähigkeit ist ein starkes Motiv für Datenproduzenten, Daten zu veröffentlichen und wird somit langfristig zu einer Verbesserung von Datenqualität und –verfügbarkeit führen.

Das Konzept stößt bei den Datenproduzenten auf breite Resonanz, dennoch wird erwartet, dass es noch Jahre brauchen wird, bevor diese neue Form der Publikation in der Wissenschaft allgemein akzeptiert wird.

Die in den Deutschen WDC entstehenden Referenzsysteme bedürfen einer Extrapolation in das globale Umfeld. Mit der Vernetzung der ICSU WDC ist ein erster Schritt für die Entstehung eines globalen Bibliotheksverbunds für wissenschaftliche Daten getan. Ein solcher Verbund wäre transdisziplinär und bietet den Vorteil, dass sämtliche Daten im Sinne des offenen Zugangs ohne Restriktionen zur Verfügung gestellt werden.

Um einerseits Langzeitsicherung wissenschaftlicher Daten zu gewährleisten andererseits die Arbeit der Datenzentren weiter zu entwickeln in Richtung auf Standards sowohl für die Bearbeitung, Archivierung und Publikation als auch die Dateninfrastruktur, d.h. Interoperabilität von Datenzentren, fehlt jedoch bislang ein gesicherter Rahmen. Die Revision des ICSU WDC Systems kann dabei nur eine konzeptionelle Vorgabe leisten. Ein Langzeitbetrieb erfordert die Absicherung durch nationale und internationale Verträge. Bislang bestehen für die WDC keine justiziablen Verpflichtungen. Dies gilt auch für PANGAEA®.

Referenzen

Berliner Erklärung über offenen Zugang zu wissenschaftlichem Wissen (2003), http://www.mpg.de/pdf/openaccess/BerlinDeclaration_dt.pdf

DFG (1998) Empfehlungen der Kommission "Selbstkontrolle in der Wissenschaft", http://www.dfg.de/aktuelles_presse/reden_stellungnahmen/download/empfehlung_wiss_praxis_0198.pdf

Diepenbroek M, Grobe H, Reinke M, Schindler U, Schlitzer R, Sieger R, Wefer G (2002) PANGAEA — an information system for environmental sciences. Computers & Geosciences 28: 1201–1210

ESF (2000) Good scientific practice in research and scholarship, http://www.esf.org/typo3conf/ext/naw_securedl/secure.php?u=0&file=fileadmin/be_user/CEO_Unit/Science_Policy/ESPB10.pdf&t=1182214053&hash=911cf3e6d783883eb9b83ee634c36d9a

Klump J, Bertelmann R, Brase J, Diepenbroek M, Grobe G, Höck H, Lautenschlager M, Schindler U, Sens I, Wächter J (2006) Data publication in the open access initiative, Data Science Journal, Vol. 5 (2006) pp.79-83, http://www.jstage.jst.go.jp/article/dsj/5/0/5_79/_article

NASA Consultative Committee for Space Data Systems (2002) Reference Model for an Open Archival Information System (OAIS) - http://public.ccsds.org/publications/archive/650x0b1.pdf

OECD (2004) Science, Technology and Innovation for the 21st Century. Meeting of the OECD Committee for Scientific and Technological Policy at Ministerial Level, 29-30 January 2004 - Final Communique, http://www.oecd.org/document/0,2340,en_2649_34487_25998799_1_1_1_1,00.html

Schindler U, Brase J, Diepenbroek M (2005) Webservices infrastructure for the registration of scientific primary data. In: Rauber A et al. (Eds.): ECDL 2005, LNCS 3652, Springer-Verlag Berlin Heidelberg: 128–138

Schindler U, Diepenbroek, M (2007) Generic Framework for Metadata Portals. Computers & Geosciences, submitted.

Sieger R, Grobe H, Diepenbroek M, Schindler U & Schlitzer R (Eds) (2005) International Collection of JGOFS - Volume 2: Integrated Data Sets (1989-2003), WDC-MARE Reports 0003 (2005), ISSN 1611 – 6577, http://www.wdc-mare.org/reports/wdc-mare_0003.pdf