

Datenpublikation im Internet

„Es gibt kein anderes wissenschaftliches Datenmodell, das mit solchen Volumina bei gleichzeitig hoher Komplexität klarkommt.“

— Dr. Hannes Grobe,
Wissenschaftlicher Mitarbeiter am
Alfred-Wegener-Institut

KEY BENEFITS

- Milliarden von Messwerten stehen über Internet-Portale für weiterführende Analysen zur Verfügung
- Jeder Datensatz ist zuverlässig auffindbar und im bibliografischen Sinn zitierfähig
- Interessenten können unterschiedlichste Aspekte kombinieren und Teilmengen extrahieren
- Ganz neue wissenschaftliche Fragestellungen sind möglich geworden

SYBASE TECHNOLOGIE

- Sybase IQ

BRANCHE

- Wissenschaft und Forschung

ÜBERSICHT

Um Daten der Klima- und Umweltforschung allgemein verfügbar und einer weiteren wissenschaftlichen Auswertung zugänglich zu machen, hat das Alfred-Wegener-Institut für Polar- und Meeresforschung (AWI) gemeinsam mit Forschungszentren an der Universität Bremen ein Informationssystem für georeferenzierte Mess- und Beobachtungswerte aufgebaut. In dem als Langzeitarchiv betriebenen System PANGAEA werden Milliarden von Messwerten aus der Erdsystemforschung gespeichert. Sie stehen der Wissenschaft über Internet-Portale für weiterführende Auswertungen und Analysen zur Verfügung. Dieses auch als Data Warehouse betriebene System hat sich inzwischen zu einer zentralen Bibliothek für eine Vielzahl geowissenschaftlicher Disziplinen entwickelt.

DIE INSTITUTIONEN

Das Alfred-Wegener-Institut für Polar- und Meeresforschung wurde 1980 in Bremerhaven gegründet. Es führt wissenschaftliche Projekte in der Arktis, Antarktis und in Meeresgebieten der gemäßigten Breiten durch. Die zur Hermann von Helmholtz-Gemeinschaft Deutscher Forschungszentren (HGF) gehörende Stiftung des öffentlichen Rechts koordiniert die Polarforschung in Deutschland und stellt die für Polarexpeditionen erforderliche Logistik zur Verfügung. Das Institut beschäftigt rund 780 Mitarbeiterinnen und Mitarbeiter. Im Zentrum für Marine Umweltwissenschaften (MARUM) und im DFG-Forschungszentrum „Ozeanränder“ (RCOM) arbeiten seit 1996 beziehungsweise 2001 der Fachbereich Geowissenschaften und andere Fachbereiche der Universität Bremen, das Alfred-Wegener-Institut, das Max-Planck-Institut für Marine Mikrobiologie in Bremen, das Zentrum für Marine Tropenökologie in Bremen sowie das Forschungsinstitut Senckenberg in Wilhelmshaven zusammen. Ziel von MARUM und RCOM ist es, die Rolle des Ozeans im System ‚Erde‘ mit modernsten geowissenschaftlichen Methoden zu entschlüsseln.

HERAUSFORDERUNG

Die immer umfangreicher werdenden wissenschaftlichen Datensätze mit ihren zum Verständnis notwendigen Metadaten werden relational in einer großen Tabelle als Teil des PANGAEA-Datenmodells gespeichert. Diese muss nach unterschiedlichsten Kriterien schnell und flexibel ausgewertet werden können.

LÖSUNG

Mit Sybase IQ besitzt PANGAEA eine technische Infrastruktur, die durch hohe Komprimierung und automatische Indizierung flexible und performante Recherchen effizient unterstützt.

ERGEBNISSE

Die exponentiell wachsenden Datenmengen werden mit dem PANGAEA-Datenmodell effizient verwaltet. Interessenten können durch die Kombination unterschiedlichster Kriterien und die gezielte Extraktion beliebiger Teilmengen aus einer umfangreichen Datenbasis ganz neue wissenschaftliche Fragestellungen formulieren.

„Angesichts der rapide wachsenden Menge von wissenschaftlichen Publikationen und Daten ist mehr Prägnanz der Information notwendig - und am prägnantesten sind nun mal strukturierte Datenmodelle mit einem freien Zugang zu ihren Inhalten.“

— Dr. Michael Diepenbroek,
Geschäftsführer WDC-MARE

Flexible Zeitreihen-Analysen aus Milliarden von Messwerten

Nährstoffe im Ozean beeinflussen das Plankton-Wachstum, das wiederum in Wechselwirkung mit dem Klima steht - somit ist die Planktonverteilung in den Weltmeeren zum Beispiel ein wichtiger Parameter in der Klimaforschung. Um festzustellen, wie sich etwa der Anteil von Nitrat und Silikat in den 1980er Jahren verändert hat, wählt Dr. Rainer Sieger, Wissenschaftlicher Mitarbeiter beim Alfred-Wegener-Institut für Polar- und Meeresforschung in Bremerhaven, am Bildschirm aus 46.000 Parametern diese beiden Typen aus. Als Zeitraum legt er 1. Januar 1980 bis 31. Dezember 1989 fest, der Vergleichbarkeit halber nur Messwerte aus einer Tiefe von 45 bis 55 Metern unter dem Meeresspiegel. Ein Knopfdruck - und Sekunden später ist aus rund 1,6 Milliarden Daten im Informationssystem PANGAEA die entsprechende „Scheibe“ selektiert. Die gewonnene Datentabelle wird mit einer speziellen Grafikanwendung aufbereitet, und Sieger kann auf einer Weltkarte die (in verschiedenen Farben angezeigte) Konzentration der gesuchten Nährstoffe in den Ozeanen studieren.


Die Klima- und Umweltforschung entwickelt immer wieder neue Fragestellungen, um die Veränderungen der Umwelt besser zu verstehen. Aber Recherchen wie diese sind gerade in der Wissenschaft noch lange nicht selbstverständlich. Zwar steht im Prinzip eine permanent wachsende Menge an Daten zur Verfügung, die teilweise auch in einer Vielzahl von Archiven weltweit gespeichert sind. Allein die Werkzeuge des Alfred-Wegener-Instituts wie das Forschungsschiff „Polarstern“ oder das luftchemische Observatorium der Neumayer-Station in der Antarktis liefern auf jeder Expedition Millionen von Daten. Sie bilden wiederum nur einen winzigen Bruchteil der Projekte zur Erdsystemforschung auf nationaler, europäischer und internationaler Ebene.

Doch genau diese Fülle steht mittlerweile einer weiterführenden wissenschaftlichen Erkenntnis eher im Wege. „Daten werden mit viel Aufwand und hohen Kosten gewonnen, sind aber häufig nicht langfristig und allgemein verfügbar“, beschreibt Dr. Hannes Grobe, Wissenschaftlicher Mitarbeiter am Alfred-Wegener-Institut, das Problem. Ein Teil von ihnen wird in Publikationen in Tabellenform angehängt, aber nur wenige sind in allgemein zugänglichen Speichern abgelegt - und dann meist in heterogenen Strukturen und proprietären Formaten, die keine systematische Auswertung erlauben. „Die Ordnung, in der sie gespeichert sind, ist primär abhängig von der Fragestellung, die hinter der Datenerfassung stand“, so Grobe. Zwar gibt es diverse internationale Standardisierungsprojekte, die aber Zeit brauchen. „Mit der zunehmenden Datenfülle seit den 80er Jahren wurden viele Daten überhaupt nicht mehr publiziert, geschweige denn abgedruckt; sie gehen damit der wissenschaftlichen Forschung verloren, da sie nicht zentral verfügbar sind“, konstatiert Dr. Michael Diepenbroek, Geschäftsführer des World Data Center for Marine Environmental Sciences (WDC-MARE). Auch WDC-MARE verwendet PANGAEA als zentrales Datenarchiv. Diepenbroeks Schlussfolgerung: „Angesichts der rapide wachsenden Menge von wissenschaftlichen Publikationen und Daten ist mehr Prägnanz der Information notwendig - und am prägnantesten sind nun mal strukturierte Datenmodelle mit einem freien Zugang zu ihren Inhalten.“

Langzeitarchiv der Erdsystemforschung

„Datenpublikation“ und zitierfähiger Datensatz heißt deshalb das Konzept, das die Verfügbarkeit von Daten verbessern soll. Entwickelt wurde es am WDC-MARE, das ebenso wie PANGAEA vom Alfred-Wegener-Institut gemeinsam mit MARUM und RCOM betrieben wird. Das Publikationssystem ist wiederum PANGAEA als öffentlich zugängliche Datenbibliothek.

Bereits im Entwicklungskonzept hat das Team die Voraussetzungen geschaffen, um PANGAEA für eine ‚Veröffentlichung‘ im bibliografischen Sinne verwenden zu können. Grundsätzlich werden die Daten mit den beschreibenden Metadaten (etwa: wer hat die Daten erhoben, um welchen Messwerttyp handelt es sich etc.) gemeinsam gespeichert. Das Datenmodell wurde so generisch und offen gehalten, dass es jederzeit um neue Parameter erweitert werden und sich damit neuen wissenschaftlichen Entwicklungen anpassen kann. Alle Daten sind georeferenziert in Zeit und Raum. Die Messpunkte sind durch drei Raumkoordinaten bestimmt (geografische Breite und Länge sowie Höhe beziehungsweise Tiefe); über die Metadaten wird außerdem erkennbar, welcher Wert wann, von wem, wie und in welchem Medium (Wasser, Luft, Eis, Sediment etc.) gemessen wurde.



Da das Datenmodell streng normalisiert ist, sind neue Messgrößen leicht zu integrieren. Soll etwa eine Planktonart als zusätzlicher Parameter aufgenommen werden, muss in der Tabelle keine weitere Spalte angelegt werden (was die Struktur verändern würde), sondern in einer neuen Zeile wird definiert, wie dieser Typ zu interpretieren ist. Über einen internen Namen wird die Referenz auf die große Messwert-Tabelle hergestellt, die den Wert selbst enthält. Damit ist die Tabelle immer gleich aufgebaut. Gerade dies ist eine wesentliche Voraussetzung für vielfältige Anfragen an ein wissenschaftliches Data Warehouse, welches die unterschiedlichsten Messwert-Typen mit einbeziehen soll.

Neben Datensätzen und Tabellen sind in PANGAEA auch binäre Objekte (BLOBs) wie Bilder oder Modelle gespeichert. Auch sie sind georeferenziert und in der Datenbank zu recherchieren. Eine URL verweist dann auf das Objekt selbst. „Mit diesem Konzept erlaubt PANGAEA die Erfassung nahezu aller in der naturwissenschaftlichen Grundlagenforschung anfallenden geografisch und zeitlich einzuordnenden Daten“, betont Grobe.

Technische Grundlage ist Sybase IQ

Die technische Grundlage dieses Konzepts ist eine spezielle analytische Datenbank, Sybase IQ. In ihr werden die validierten und publizierten (und danach unveränderlichen) Daten gespeichert. Dass PANGAEA eine Vielzahl von Attributen und Messwerten in einer einzigen großen Tabelle zur Verfügung stellen kann, basiert nicht zuletzt auf der hohen Komprimierungsfähigkeit (auf ein Sechstel bis ein Neuntel der Rohdatenmenge) sowie dem Indizierungsprinzip dieser Datenbank.

Bei klassischen relationalen Datenbanken müssen vorab vielfältige Indizes angelegt werden, um gezielte Recherchen durchführen zu können. Dadurch wird das Datenvolumen erheblich aufgebläht; eine Beschränkung reduziert wiederum die Abfrageflexibilität. Bei Sybase IQ sind die Daten nicht in Reihen, sondern als Spalten organisiert, wobei automatisch jedes Feld als Index dienen kann. Damit sind Recherchen sehr flexibel: Egal welche Fragestellung kommt - die Datenbank bietet immer Zugriffshilfen. Da nicht die gesamte Zeile, sondern nur noch der ausgewählte Wert in der entsprechenden Spalte gelesen wird, sind Abfragen auch bei extrem großen Datenmengen sehr schnell.


Bibliothekskonzept für die Datenpublikation genutzt

Um Wissenschaftler zu motivieren, mehr Daten zu archivieren und zu publizieren, sollten diese im bibliographischen Sinne zitierfähig sein. Um dies zu ermöglichen, bildet PANGAEA den etablierten Arbeitsfluss Autor - Verlag - Bibliothek ab. Zur Unterstützung einer zuverlässigen Verfügbarkeit wurde ein Code für die Kennzeichnung wissenschaftlicher Publikationen (Digital Object Identifier - DOI) auf die Datenwelt übertragen. Seit 2004 erhält jeder veröffentlichte Datensatz eine solche Referenznummer, die von der TIB Hannover (der größten wissenschaftlich-technischen Bibliothek der Welt) vergeben wird. Damit ist er zuverlässig auffindbar und im bibliografischen Sinn zitierfähig. PANGAEA ist mit weiteren Partnern Initiator der Anwendung des DOI-Systems auf die Publikation von Daten.

Ein Autor wird jetzt bei Nutzung seiner Daten in wissenschaftlichen Publikationen sachgerecht zitiert. Der Verwender des Zitats wiederum hat die Gewissheit, dass es sich um Daten handelt, die von einem Autor validiert veröffentlicht wurden - und der somit auch für die Qualität verantwortlich ist. Mit der DOI kann ein Datensatz jederzeit über das Internet direkt aufgerufen werden. Da es sich um einen dauerhaften Link handelt, wird die physische Adresse der Daten auch dann gefunden, wenn sich die Lokation zwischenzeitlich geändert hat.

Effiziente Zeitreihenanalysen

Die Daten in PANGAEA werden in einem Datenbank-Server (Sybase IQ) verwaltet. Interessenten können darauf über Clients oder Internet-Portale zugreifen (von denen das Alfred-Wegener-Institut selbst fachspezifische aufbaut). Unter der URL <http://www.pangaea.de> stehen alle Daten über eine allgemeine Suchmaschine mit Google-ähnlicher Syntax der wissenschaftlichen Gemeinschaft zur Verfügung.



Außerdem entwickeln Diepenbroek und der Diplomphysiker Uwe Schindler von der Universität Bremen derzeit ein universelles Zugriffs-Werkzeug, mit dem beliebige Teilmengen aus dem Datenpool extrahiert werden können. „Damit können Nutzer, die mehr als 40.000 Messgrößen einzeln oder in beliebiger Kombination benötigen, diese in jeder raumzeitlichen Menge mit der gewünschten Filterung herausziehen - mit Antwortzeiten im Sekundenbereich“, beschreibt Diepenbroek den entscheidenden Fortschritt. Wissenschaftler können auf diese Weise individuell zusammengestellte Teilmengen zum Beispiel von Zeitreihen betrachten. Auch dafür bietet Sybase IQ durch eine effiziente Möglichkeit zur Indizierung von Zeit-Informationen eine technische Unterstützung.

Ein weiteres Beispiel demonstriert Dr. Christian Schäfer-Neth, Wissenschaftlicher Angestellter im Alfred-Wegener-Institut, mit der Analyse von Daten der Neumayer-Station. Das luftchemische Observatorium kann in der sehr sauberen Luft der Antarktis unbeeinflusst von lokalen menschlichen Einflüssen langfristige Entwicklungen besonders genau verfolgen. Gemessen werden in bestimmten Abständen (Minuten und Stunden) unter anderem Ozon, Partikel, Treibhausgase, Spurengase, Aerosole und Trübung. Neben den aktuellen Luftwerten lassen sich durch Bohrungen im Eis Entwicklungen in der Erdatmosphäre der letzten hunderttausend Jahre zurückverfolgen. Alle gewonnenen Werte werden in PANGAEA archiviert und erlauben so die Verschneidung auch extrem langer Zeitreihenbetrachtungen aus historischen und aktuellen Werten.

Um beispielsweise die Ozon- und Partikelkonzentration in den 1990er Jahren festzustellen, gibt Schäfer-Neth diese beiden Parameter zusammen mit der Zeitdefinition (1/1990 bis 12/1999) ein – und erhält in Sekundenschnelle aus rund 11 Millionen Luftchemie-Daten ein Diagramm mit den gewünschten Mittelwerten.

Zentralarchiv für wissenschaftliche Daten

Zeitreihen, ozeanografische und seismische Profile, Sedimentprofile, geologische Karten - die Daten aus PANGAEA lassen sich nahezu unbegrenzt auswerten. Und Fülle wie Vielfalt wachsen ständig. Hat PANGAEA ursprünglich mit den Daten des Alfred-Wegener-Institut begonnen, hat es sich mittlerweile zu einem zentralen Archiv für viele wissenschaftliche Daten entwickelt. Durch das Datenmodell mit der Möglichkeit zur universellen Recherche, die Masse der mittlerweile gespeicherten Daten und das DOI-Konzept entfaltet es eine Sogwirkung auf externe Datensammlungen, die teilweise Fördermittel erhalten, um ihre Daten in PANGAEA zu archivieren. Beispielsweise wurde inzwischen eines der weltweit umfangreichsten Archive für Baumringe eingebracht – ein wichtiger Beitrag für die Paläoklimaforschung, die erdgeschichtliche Untersuchung des Weltklimas. Bei zahlreichen Projekten auf nationaler, europäischer und internationaler Ebene wird PANGAEA von vornherein mit dem Datenmanagement betraut.

EingabeprozEDUREN und technische Routinen unterstützen den Datenimport. Um die Konsistenz sicherzustellen, obliegt die Verantwortung für die technische Qualität der Daten bei jedem Projekt einem Kurator. Erst nachdem er einen Datensatz freigegeben hat, wird dieser mit einer DOI-Nummer versehen und publiziert. Zukünftig soll auch ein Peer-Review die wissenschaftliche Qualität der Daten unterstützen. Auch die Qualitätssicherung selbst konnte durch die Bündelung einer derart großen Datenmenge in einem einzigen Data Warehouse optimiert werden. Diepenbroek: „Wenn wir beispielsweise einen Datensatz mit 5.000 Werten eingeben und darunter ist ein Ausreißer, kann sich dahinter ein Messfehler oder ein extremes Ereignis verbergen. Das können wir nur kontrollieren, wenn wir einen hinreichend großen Datenpool haben, in dem ähnliche Phänomene abgebildet sind.“

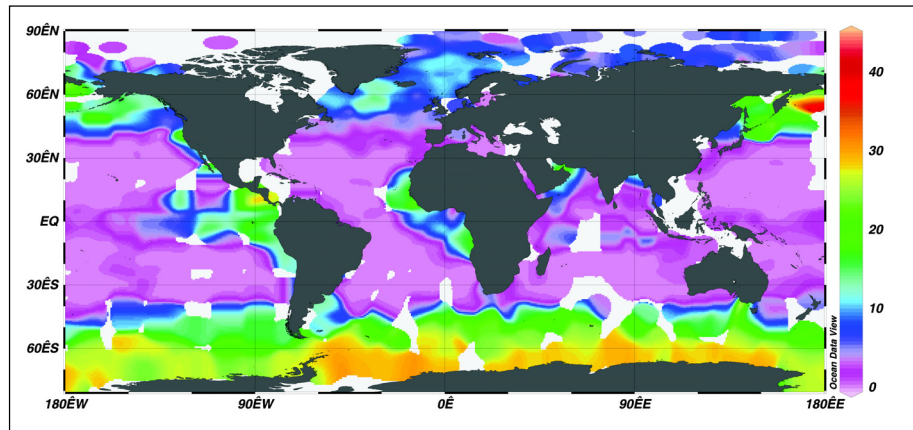
Der Inhalt von PANGAEA ist in den vergangenen Jahren exponentiell gewachsen; seit 1996 hat sich die Datenmenge jedes Jahr verdoppelt. Im November 2006 waren 136 Forschungsprojekte mit 470.000 Datensätzen und 1,7 Milliarden Datenpunkten aufgenommen, wodurch ein Datenvolumen von 1,2 Terabyte erreicht wurde. Und die Datenmenge wächst permanent weiter; so liefern einige Beobachtungsstationen neue Werte im Minutentakt. Die im Aufbau befindlichen Sensornetzwerke werden aus Ozeanografie, Meteorologie und Geophysik den Datenfluss um nochmals ein bis zwei Größenordnungen erhöhen.

„Indem wir eine solche Menge an Daten zusammen mit den Metadaten in einem zentralen System speichern, haben wir eine enorme Effizienzsteigerung erreicht“, resümiert Grobe. „Es gibt kein anderes wissenschaftliches Datenmodell, das mit solchen Volumina bei gleichzeitig hoher Komplexität klarkommt. Da unterschiedlichste Aspekte in einer einzigen Abfrage verknüpft werden können, sind damit ganz neue wissenschaftliche Fragestellungen möglich. Wir können so wichtige Unterstützung bei der Erforschung unserer Erde leisten.“

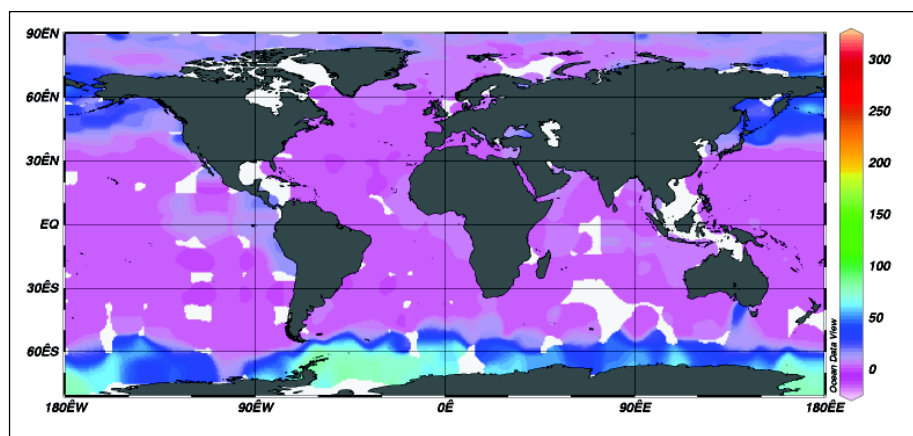
Internet-Links

www.pangaea.de, www.awi.de, www.wdc-mare.org, www.rcom.marum.de, www.std-doi.de

Weltkarte mit der Konzentration von Nitrat und Silikat in den Ozeanen auf Basis einer PANGAEA-Analyse



Nitrate [$\mu\text{mol/l}$] @ Depth, water [m] = 50 m



Silicate [$\mu\text{mol/l}$] @ Depth, water [m] = 50 m

Sybase IQ

Sybase IQ ist speziell auf die Analyse extrem großer Datenmengen zugeschnitten. Es handelt sich um eine voll relationale, SQL-fähige Datenbank, die intern eine patentierte Art der Speicherung und Verwaltung anwendet (vertikale Partitionierung). Dabei werden die Daten nicht in Zeilen, sondern in Spalten organisiert. Jedes Feld kann direkt als Abfrageschlüssel dienen; deshalb müssen keine traditionellen Indizes definiert werden. Bei Abfragen muss nicht die gesamte Zeile, sondern nur noch der ausgewählte Wert in der entsprechenden Spalte gelesen werden. Die zu verarbeitenden Datenmengen werden minimiert und die Zahl der I/O-Operationen wird bis zu 90 Prozent reduziert. Außerdem wendet Sybase das Prinzip des Bitwise Indexing an, bei dem, wo immer möglich, die Datenelemente einer Spalte in eine Bitmaske transformiert werden. Durch diese Architekturprinzipien kann die Abfragegeschwindigkeit um das Zehn- bis Hundertfache gesteigert und die Datenmenge auf ein Fünftel bis ein Neuntel komprimiert werden.