An OAI Framework for biodiversity and contextual content: „PlanktonNet“ as pilot study

Ana Macario and Bastian Onken
Ana.Macario@awi.de, Bastian.Onken@awi.de

Data and Computing Centre of
Alfred Wegener Institute for Polar and Marine Research, Germany

Alfred-Wegener-Institut
für Polar- und Meeresforschung
in der Helmholtz-Gemeinschaft

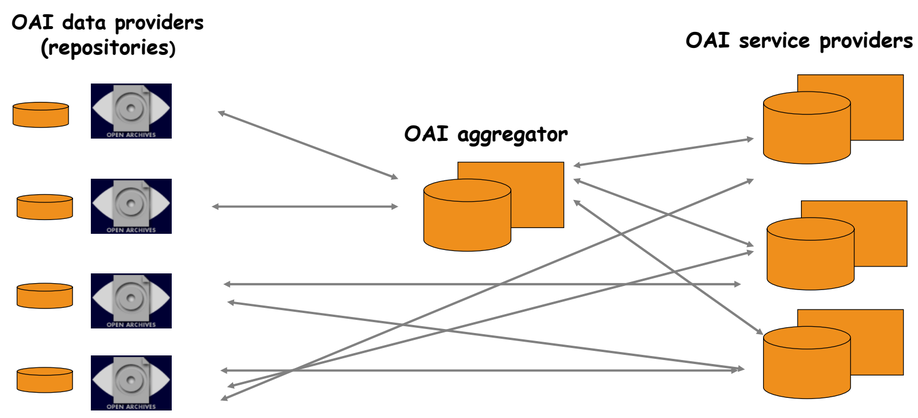
Sixth Framework Programme

Abstract

Ocean Biodiversity Informatics Conference 2007
International Conference on Biodiversity Data Management
Bedford Institute of Oceanography, Dartmouth Nova Scotia, Canada
October 2-4, 2007

Digital objects in the field of earth and biological sciences are known to be often compound and complex. If one takes biodiversity as example, an exhaustively long list of information systems can be found on-line. These systems often contain valuable and historically relevant information gathered over several decades using distinct archival resources, data models and transport protocols.

In order to assure long term preservation of this distributed and not yet networked mass of information on world's biota, we need to create an abstract interoperability layer in which digital objects sharing the same data model are aggregated so as to allow for information preservation, transformation, re-use and exchange . Using panFMP and FEDORA technologies, we propose a strategy for creating an information network for biodiversity and related contextual content (e.g., oceanographic data). A prototype for the proposed information network was built upon existing PlanktonNet data providers, publication repositories and environmental data archived in WDC-MARE/PANGAEA. Because XML schemas for metadata description and for expressing relationships among digital objects are available, interoperability with other federated networks will be assured. In addition, a panFMP front-end customized specifically for PlanktonNet is presented as metadata portal.



Why OAI?

OAI offers a low barrier **data provider and service provider** framework with no constraint on data archival architecture (RDBMS, XML, etc)

OAI accomodates any metadata schema in addition to Dublin Core. **OAI-PMH** is a widely deployed protocol standard for harvesting metadata for all types of objects. Protocol specification for resource harvesting (for object re-use and exchange purposes) will be soon available as result of the **ORE** initiative

OAI allows for **incremental and selective harvesting** of individual collections/sets

OAI allows for seamless access to **distributed repositories** and thus flexibility in the **re-use** of objects and **discovery** of content in different contexts. It bridges the gap between biodiversity archives and other repositories (e.g. publications, geological and environmental data repositories, etc). In addition, OAI records are **cite-ready** and are harvested by **Google Scholar** (OAI-PMH is part of Google's sitemap protocol)

PlanktonNet components

Data providers

PlanktonNet OAI compliant data providers (PlanktonNet@AWI, PlanktonNet@Lisbon, PlanktonNet@Roscoff, PlanktonNet@Israel)
total no. of records: 4,278

Environmental data (WDC-MARE/PANGAEA)
total no. of records: 562,916

Publications
total no. of records: 10,437

Aggregator panFMP

panFMP (PANGAEA framework for metadata portal) is generic and flexible harvester powered by Apache Lucene indexing and searching engine

FEDORA repository

Preservation of digital objects; semantic and dissemination services planned in the future

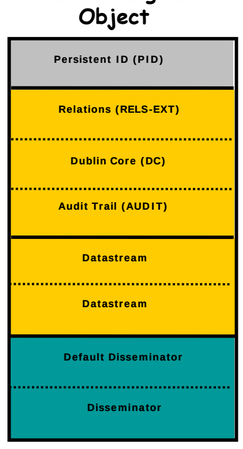
Why FEDORA?

FEDORA offers a **scalable** open access **repository** framework compliant with international standards (XML storage, OAI-PMH, SOAP/REST web services, etc). The flexible and extensible digital object model behind FEDORA allows any metadata description schema and **integrity checking** (schema validation).

FEDORA assures object **preservation** though **content versionning**, and **control access** at object and collection levels.

FEDORA architecture includes a generic **RDF-based relationship model** that represents relationships among objects and their components.

FEDORA's ability to **distribute load and object storage** among several IR instances („Virtual Repository“ concept) in a federated environmental together with ist **semantic services** are key features for a succesful network of biological information systems.



RDF-based relationship ontology

Collection-related: isMemberofCollection, isMemberofAsset, isRelated, HasImage

Branding-related: AggregatedBy, DescribedBy, CertifiedBy

Service-oriented disseminators

Metadata-crosswalks (getDarwinCore2, getOBIS, getABCD, ...)

Image transformation and annotation services

LSID services (TDWG/GUID)

uBio Taxonomic „intelligence“ services

OGC/GML and KML services (OpenGIS, GoogleEarth, GeoRSS-GML)

RSS Feeds services (GeoRSS)

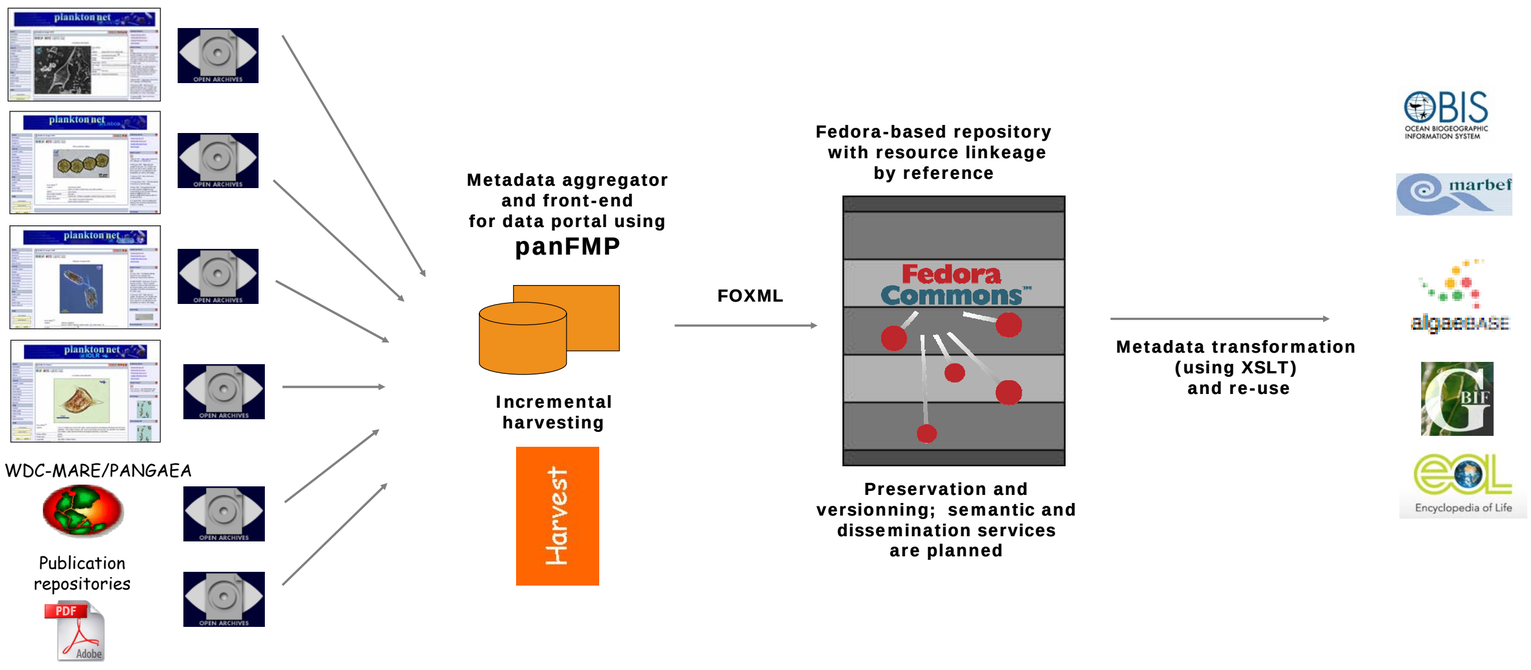
Why panFMP?

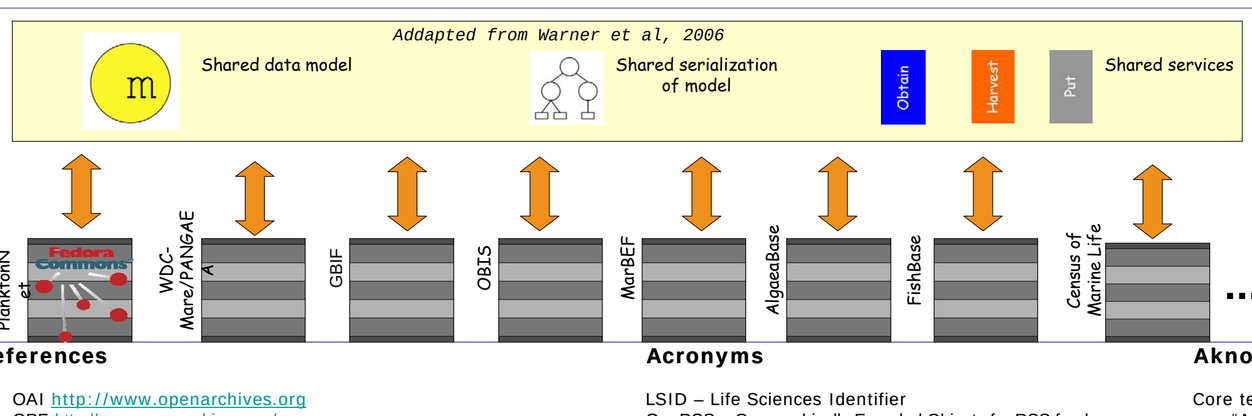
panFMP offers a generic and flexible framework for building metadata portals based on **Apache Lucene** indexing and search technology.

panFMP works as **aggregator** for PlanktonNet and contextual content: we are able to currently harvest OAI-PMH compliant PlanktonNet data providers, several well-established institutional repositories for publications and the world centre for geological and environmental data WDC-MARE/PANGAEA. A web-based front-end for panFMP customized specifically for the needs of PlanktonNet project was developed [http://data.planktonnet.eu]

panFMP supports **any XML format** : data providers can be harvested with any of the commonly used protocols (currently tested with OAI-PMH and OGCCS) and metadata description formats (e.g., Dublin Core, ISO 19115, Darwin Core2, ABCD, OBIS...). Because the harvested metadata are stored in separate indexes, these can be combined accordingly to serve distinct purposes of individual portals.

The harvested indexes are exposed via **SOAP** web services through a java API. Long-term preservation, versionning and ACL issues can be handled by archiving the harvested metadata in a repository framework of choice (e.g., FEDORA)





Interoperability lessons from „Pathways“

A long list of biological information systems based on heterogeneous data models/interfaces and using distinct metadata description and transport protocols are currently available.

In order to offer valuable service-oriented gateways targeted at specific projects (e.g. PlanktonNet data portal), simultaneous access to as many as possible repositories is wished. True repository interoperability can only be accomplished by agreeing in an **interoperability layer** in which the **data model (including granularity) and services** are commonly shared across repositories. Further details can be read on NSDL's Pathway project

By using relationshop **ontology** concepts one can express valuable **relationship metadata** across repositories and preserve institutional **branding** when applicable. This aspect might be in particular relevant for biodiversity content given the wide range of **curation quality**.

References

1. OAI http://www.openarchives.org
2. ORE http://www.openarchives.org/ore
3. Fedora Commons http://fedora-commons.org
4. panFMP, Pangaea Framework for Metadata Portal [Schindler, U, Diepenbroek, M, 2007. Generic Framework for Metadata Portals. Computers & Geosciences, submitted], visit http://www.panfmp.org;
5. Pangaea http://www.pangaea.de
6. Warner et al (2006) Pathways: Augmenting interoperability across scholarly repositories http://arxiv.org/abs/cs/0610031
7. PlanktonNet site http://www.planktonnet.eu
8. PlanktonNet data portal http://data.planktonnet.eu
9. Taxonomic databases working group, subgroup GUID http://wiki.tdwg.org/twiki/bin/view/GUID

Acronyms

LSID – Life Sciences Identifier
GeoRSS - Geographically Encoded Objects for RSS feeds
GML – Geography Markup Language
GUID – Glocal Unique Identifiers
OAI - Open Archives Inititative
OAI-PMH - Open Archives Inititative Protocol for metadata harvesting
OGC – Open Geospatial Consortium
OGCCS - Open Geospatial Consortium Catalogue Service
ORE - Object resource and exchange
PANGAEA – Publishing Network for Geoscientific and Environmental Data
RDF – Resource Description Framework
RSS – Really Simple Syndication
TDWG – Taxonomic Databases Working Group

Aknowledgments

Core terms and graphic conventions according to guidelines of the “Augmenting interoperability across scholarly repositories,” meeting (09/2006) have been used in this poster [http://msc.mellon.org/Meetings/Interop/]

The authors thank Carl Lagoze and Sandy Payette from FEDORA team for initial discussions on content models and Uwe Schindler for producing the open source aggregator framework as part of his dissertation thesis.

This work was funded by an award from the Sixth EU Framework Programme („PlanktonNet“) and Alfred Wegener Institute for Polar and Marine Research, Germany.

Cite us

Macario, A. and Onken, B (2006). An OAI framework for biodiversity and contextual content: PlanktonNet as pilot study. Ocean Biodiversity Informatics International Conference - OBI'07, Bedford Institute of Oceanography, 02-04 October 2007, Nova Scotia, Canada [hdl: 10013/epic.27718]