



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



Text, Data and People – How to Represent Earth System Science

Hans Pfeiffenberger

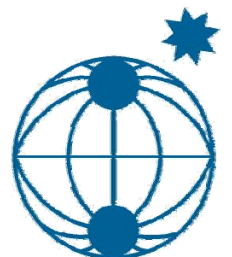
Ana Macario

Alfred Wegener Institut, Bremerhaven

1

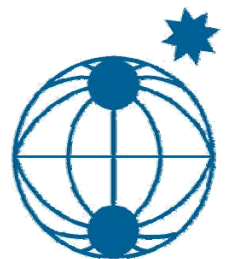
Hans Pfeiffenberger, Ana Macario, Alfred Wegener Institut, Helmholtz Association

OAI4 CERN 2005-10-20



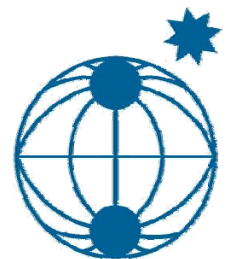
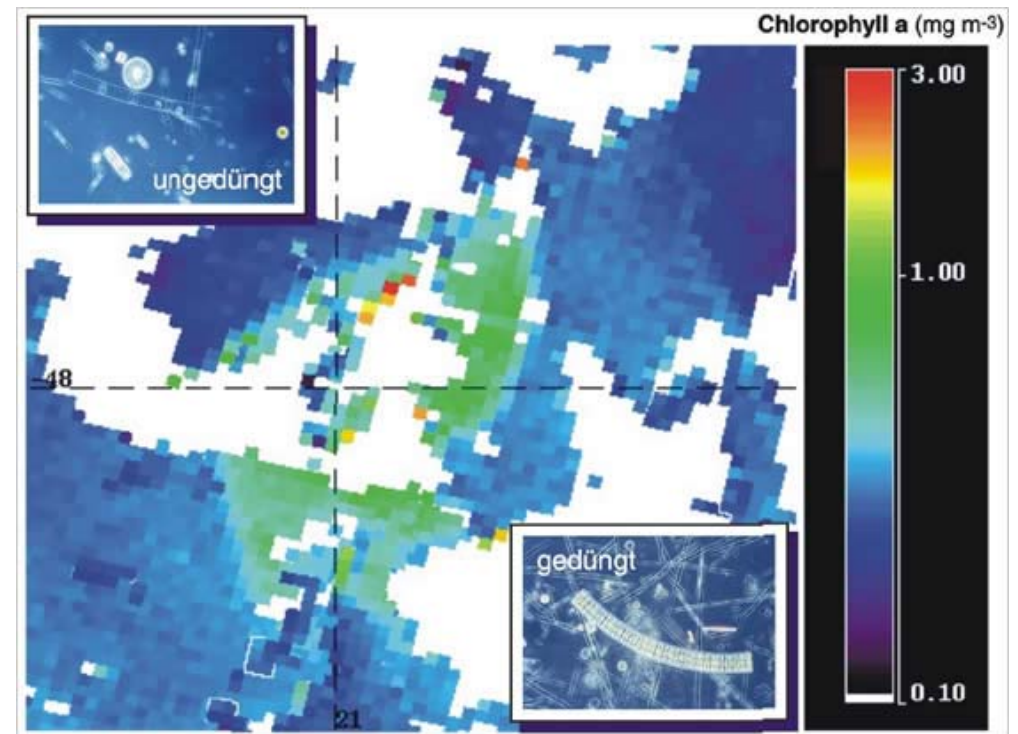
Introduction

- *Earth System Science (ESS) is an **interdisciplinary and global collaboration***
- *ESS output is heavily data-centric*
 - **data come from observations**
 - **and simulation (“in silico” experiments)**
- *ESS work is organized around*
 - **expeditions or campaigns** and
 - **coupled models of earth’s sub-systems**
- *Logistics and system **cost are extremely high***
 - **one ship may cost up to 500 G€**
 - **“Earth Simulator”, the fastest computer 2 years ago**
- *ESS data potentially are of extreme **long term value***



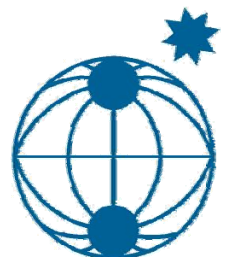
An important, typical Experiment

- *EISENEX / EIFEX* : Conducted during two expeditions of “Polarstern”, with a 4 year pause
- *EIFEX (2004)*:
 - 54 scientists (and students) from
 - 14 institutes and 3 companies from
 - 7 European countries and South Africa
 - Oceanographers
 - Biologists
 - Chemists.....
- “*Biogeochemistry*”



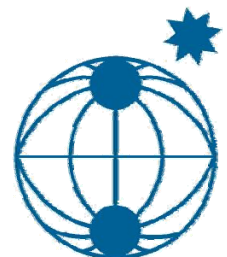
Collaboration's data needs

- *Need to work from a common understanding of what is known about the subject*
- *Need to plan expeditions and coordinate with ships' operators general plan (5 or more years in advance)*
- *Need to coordinate instrument design, operation and interfacing before ships departure*
- *Meet aboard , sail and work 8 weeks or so*
- *Do evaluation, when at the home institute, exchanging their particular results.*
- *Publish text; PhD students dump the data somewhere, if nobody watches, or keep it "private"*



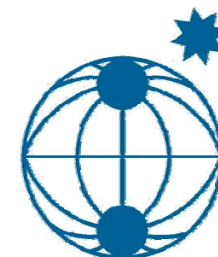
Data Publishing

- *There is reason enough to thoroughly publish data:*
 - Potential **reuse** in many more contexts than foreseen
 - Enable **peer reviewers** to have a critical look at data quality
- *Problem: Metadata*
 - ISO 19115 is a metadata standard (with ~1000 attributes) for georeferenced data
 - Almost **no producer of data knows how to form ISO 19115** for his/her data (nor wishes to know)
- *There is no reward system (like: number of peer reviewed papers) in place to stimulate individuals*
 - There should be a solution for well curated datasets and databases



Data Management

- *Metadata needed even on “work in progress”- or auxiliary datasets,*
 - both need to be “archived”, or managed
 - Even if they may never achieve a level of “published” data
 - They need to be **available to a distributed project group during their project**, long before publication
- *There are too many datasets to **produce** correct and complete ISO 19115 metadata “**manually**”*
 - Find ways to produce ISO by each instrument at the time of data creation, **automatically**
 - Use **context or relationship instead of descriptive metadata**



Relating all relevant Objects

...but for
AWI
expeditions
only, today



eXPEDITION
version 1.0

Login[Admin]

R.V. "[Polarstern](#)" - Expedition year: 2004

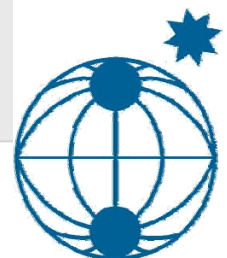
Research vessels: Polarstern

Operation: Alfred-Wegener-Institut [Germany]
Ice class: GL/ARC 3, built 1982
Length/Beam: 117.91 m/25.07 m
Purpose: Marine Science, logistic, re-supply
General info: [Current expedition](#) | [Current weather](#) | [Long-term schedule](#) | [On-line library](#)
Cruise participation: [Call for Proposals](#) | [Proposal guidelines](#) | [Cruise-related forms](#)
Instrumentation: [All](#) | [Physical oceanography](#)
Data: [Meteorology](#) | [Oceanography](#) | [PANGAEA](#) | [PODAS](#)
"Polarstern Abstracts": [All](#) | [Submit new abstract](#)
Additional Publications: [Field reports](#) | [PhD Thesis](#)
Presentations: [Invited talks](#) | [Talks](#) | [Posters](#)

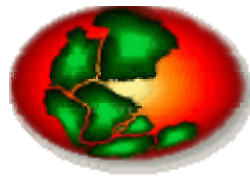
Where is "Polarstern"?

[See also] [Print schedule]

Expedition	Date Port	Region Research	Weekly reports	Data AWI Publications	Details
ANT-XXI/3 Coordinator: Pörtner, H. Chief scientist: Smetacek, V.	21.01.2004 - 25.03.2004 Capetown - Capetown [Map]	Atlantic/Indian Ocean, Polar frontal zone Biology, EIFEX [Press release]	1. Report 2. Report 3. Report 4. Report 5. Report 6. Report 7. Report 8. Report 9. Report	Meteorology PANGAEA: Stations PANGAEA: Datasets [Note: Datasets for recent cruises may not yet be available] "Polarstern Abstracts" Bathmann [2004] Bathmann [2005]a Bathmann [2005]b Bathmann [2005]c Leach et al [2004] Sachs et al [2005]	



Current PANGAEA relationship encoding



← Resource

Dataset-to-Publication
relationship metadata
should be expressed in RDF/XML
and placed in the
"Relations datastream"

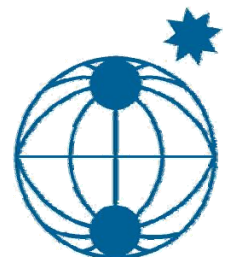
Identifiers needed (in addition to locators)

Dublin Core

Descriptive
metadata

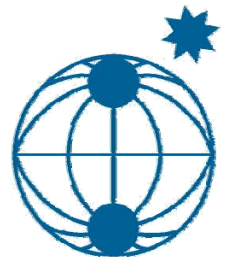
Metadata
<source>
for content

<relation>
for
relation(s)



Goals

- Transfer concepts and content *from “homegrown”, internal repositories to federations of standards-based IRs around the world*
- Harvest (f.e.) Polarstern-expedition related text and data from all *IRs of participants*
- Display / sort / analyze / rank the *maze of material* through all meaningful criteria
- Find *key networks* of people, projects, text,.....



Home

Welcome to the **Helmholtz Knowledge Management Archives for "Earth and Environment"**
Search for relevant open-access databases, associated publications and personal portfolio

 [Advanced Search](#) |  [Browse Archives](#)

Search:

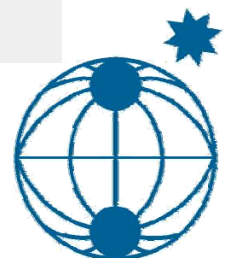
in

-- All Collections --

- All Collections --
- DataSets
- Personal Portfolio
- All Publications --
- Publication: articles
- Publication: books
- Publication: inbooks
- Publication: conference papers
- Publication: field reports
- All Events --
- Events: talks
- Events: invited talks
- Events: posters
- All Technology transfers--
- Technology transfer: patente
- Technology transfer: trademarks
- Technology transfer: utility model
- All Thesis --
- Thesis: bachelors
- Thesis: diploms

[Home](#) | [Search](#) | [Archives](#) |

© 2003-2004 Powered by
Public Knowledge Project



Welcome to **Helmholtz Web Services** for primary data, publications and personal portfolio.

Search:

Select Repository:

Fedora at AWI

Pangaea

Search

Results 1-5 for '**macario**':

[Show Results 6-10 ->>](#)

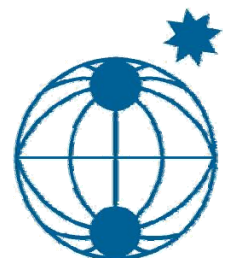
- 1** [Personal Homepage of Dr. Ana Macario](#) [text, people]
(2005) Ana Macario
- 2** [A Discovery Service for Knowledge Related to Research Platforms](#) [event, international talk]
(2004) Macario, A.; Pfeiffenberger, H.; Reinke, M.
- 3** [Portal for Earth Sciences in Polar Regions](#) [event, international talk]
(2003) König-Langlo, G.; Macario, A.; Olbers, D.; Pfeiffenberger, H.; Reinke, M.; Thiede, J.
- 4** [Research platforms in polar regions - a portal approach](#) [event, international talk]
(2003) König-Langlo, G.; Macario, A.; Olbers, D.; Pfeiffenberger, H.; Reinke, M.; Thiede, J.
- 5** [An homogeneous Directory of People, Publications, and other Resources as a means for IT-based Knowledge Management in Science](#) [event, invited national talk]
(2000) Macario, A.; Pfeiffenberger, H.

Fedora at AWI Response Time: **0.144s**, 5 Results

Pangaea Response Time: **0.374s**, 0 Results

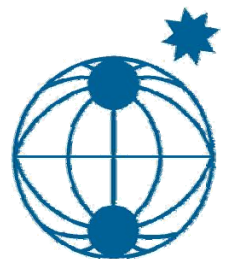
[Show Results 6-10 ->>](#)

Script Time: **0.399s**



Types of Object In the order of appearance (1)

- *(Institutions)*
- *Person*
 - represented by splash page (Personal home page)
 - uid: eduPersonPrimaryName
 - primary encoding: eduPerson schema
- *(informal group)*
- *Project*
 - represented by splash page (Project home page)
 - uid: maybe a specific encoding of the funders' project number
 - primary encoding: eduPerson/eduOrg schema
- *Expedition, Campaign:*
 - represented by splash page (Expedition home page)
 - treat it as a project, generate project number from expedition identifier
 - primary encoding: eduPerson/eduOrg schema



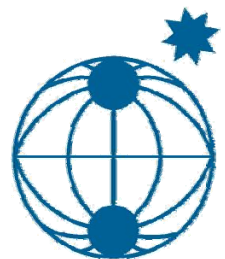
Types of Object In the order of appearance (2)

■ *Datasets*

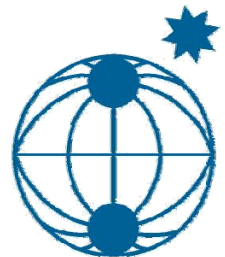
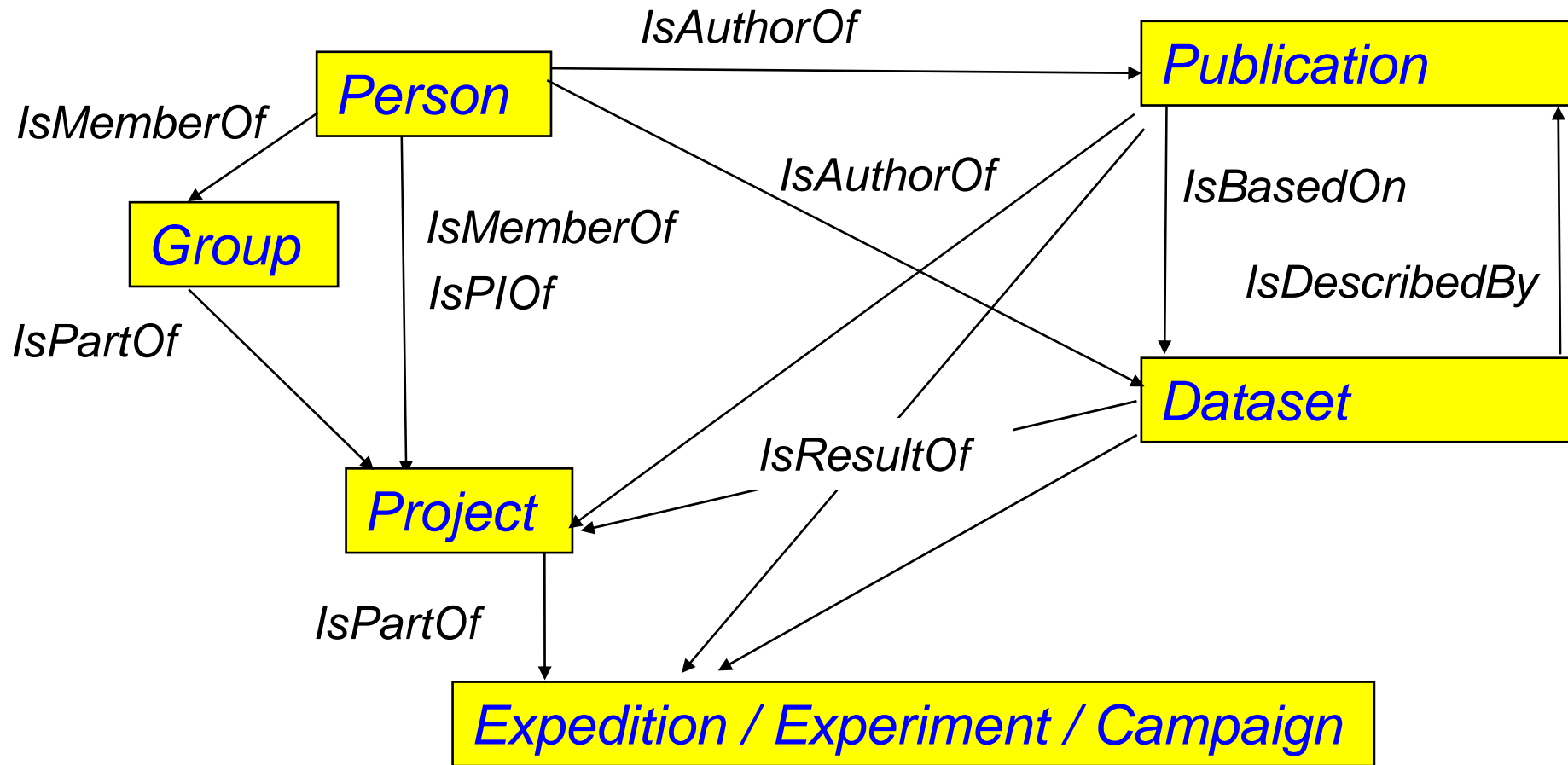
- represented by splash page
- uid: maybe the same kind as publications
- primary encoding: Community specific (f.e.: ISO 19115)

■ *Publications*

- represented by splash page containing
 - *abstract, etc.*
 - *pointer to article at publishers site*
 - *pointer to article at IR*
 - *publisher's word about what is the "original", etc.*
- uid: DOI, permanent URL, etc.
- primary encoding: repository's (proprietary) format (f.e.: Fedora's , it must be possible to map this in an unambiguous way to METS, MPEG21-DIDL,...

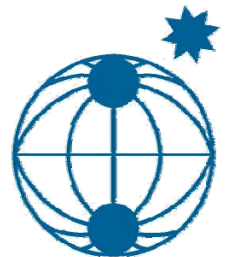


Object relationships (tentative)



Conclusion 1 – Text with Data

- *(Text-)Publications and related primary data have to be cross-referenced*
 - **We need ontology and schema designs to express the relationships (to solve reuse/aggregation problem)**
- *Extensive descriptive metadata (f.e. ISO19115) are useful only to big repositories of well curated datasets with similar content*
- *The full text of publications (and its relation to datasets) may be the best “metadata” for the datasets you will get*
 - **Primary hit in a (Google-like) search may be a publication, which refers to primary data**



Conclusions 2 - Full Relation Network

- *Service providers should make use of **network of all relevant objects** - people, projects, ... datasets, text*
 - harvest relationship metadata
 - harvest descriptive metadata (Dublin Core quality)
 - enable new search paradigms
- *Data providers need to expose the relationship of objects*
 - will require a “complex” metadata format
 - will require an ontology for relationships
 - will require unique identifiers for people etc. (from eduPerson schema , ~ email address)
 - introduce **identifiers for projects and “experiments”**

