GES 2007 German e-Science

Available online at http://www.ges2007.de

# A Network of Text, Data and People for the Earth System Sciences

H.Pfeiffenberger, A.Macario, H.Grobe

Alfred Wegener Institute for Polar and Marine Research,
Am Handelshafen 12, 27570 Bremerhaven, Germany
*Email:* {Hans.Pfeiffenberger, Ana.Macario, Hannes.Grobe}@awi.de
*phone:* (+49 471) 4831 1305, *fax:* (+49 471) 4831 1590

## Abstract

Earth System Science is an outstanding example of a field of research which yields important results especially when conducted in multidisciplinary and global cooperation. The International Polar Year and its expected legacy are used as an example to illustrate this assertion and the financial and intellectual expense invested. It follows that any effort to make more out of the globally distributed – if not fragmented – results and to network the knowledge gained would be valuable indeed. An experimental implementation of such a system, connecting journal articles to datasets, expeditions and researchers involved, is introduced. Some developments necessary to implement comparable and more powerful systems on a global scale are discussed.

## 1 Introduction

The “hard” topics of Earth System Science (ESS) span climate research, geo- and biosphere and include modelling as well as observations and experiments in the laboratory and in situ, all around the world from deep sea to the high atmosphere. Noteworthy, research into the human dimensions – social and economic- is included in the definition of ESS as well as in recent research programs like the International Polar Year 2007-2008 (IPY).

To establish the scope of eScience tools required to enhance ESS, in section 2 we describe the IPY as a major current activity.

Section 3 describes eXpedition, a portal at the Alfred Wegener Institute (AWI) which connects and relates articles, datasets, scientists involved and general descriptions of objectives and activities of AWI expeditions.

Section 4 discusses necessary conditions and developments for future systems, similar and beyond eXpedition, which need to operate on a global scale on distributed repositories and directories.

## 2 The International Polar Year 2007-2008

On 1 March 2007 the IPY celebrated its official opening ceremony and Global Launch event. The IPY is a co-ordinated program of earth sciences, economic and sociological research into the arctic and antarctic regions. The research themes comprise “atmosphere, ice, land, oceans, people and space”. One of the major objectives of IPY is to leave a data legacy, also called a snapshot (of a significant part) of the planet Earth, which “will provide a crucial benchmark for detecting and understanding change in comparison with past and future data sets"[1].

Consequently, the program involves at least 50.000 participants from 63 nations. On top of this extreme use of intellectual capital, the program makes use of very expensive research infrastructure, as for example fleets of satellites, ships, airplanes and heavy logistics. The overall expenditure is estimated to be upwards of one billion ($10^9$) Euro. The program was initiated and organized under the auspices of the International Council for Science (ICSU) and the World Meteorological Organization (WMO); its funding derives mostly from ongoing and additional national research programs and projects.

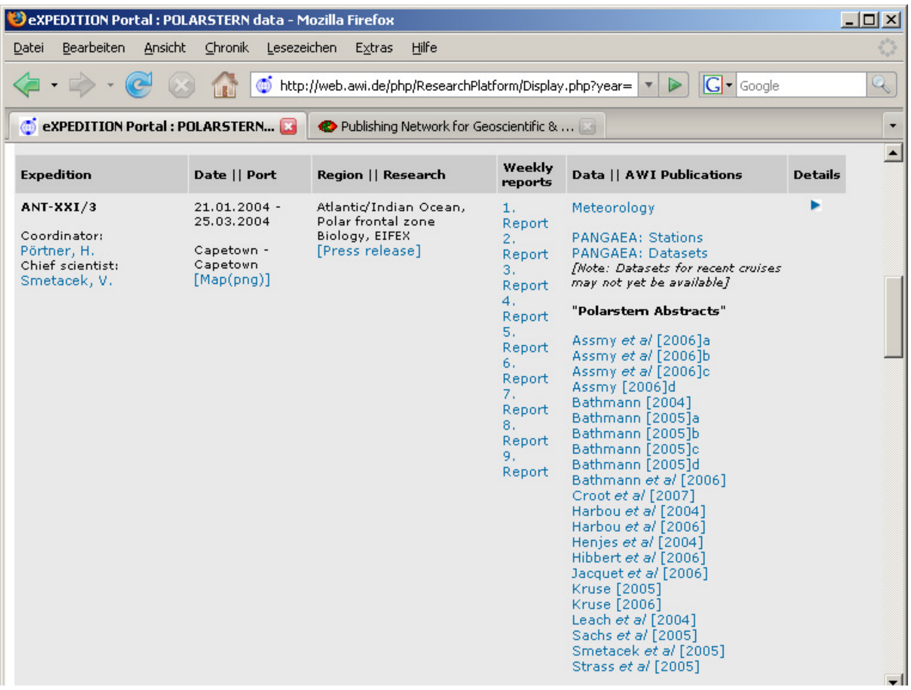
The IPY "represents one of the most ambitious coordinated international science programmes ever attempted." [2] "It aims to leave a legacy of new or enhanced observational systems, facilities, infrastructure, numerical Earth simulators and research networks, as well as an unprecedented degree of access to the data and information it will generate" [3]. Specifically, "in order to meet it's objectives of interdisciplinary and international collaboration and to ensure a lasting legacy, IPY is committed to ensuring full, free, and open access to IPY data as described in the IPY data policy" [4]. In fact, the IPY Joint Committee rephrased, in this data policy document [5], the general *commitment* to a detailed *requirement* on projects' practises.

While the most recent and comprehensive "Scope" document of IPY [2] devotes a main chapter to "IPY Data and Information Management", it is near completely geared towards the preservation and integration of *data*. Missing, conspicuously, is any explicit mention of classical publications, e.g. journal articles (Although one of the 220 IPY project, no.51, [2] is concerned with collecting a bibliography of IPY and IPY related references). This centre of attention could be interpreted in two ways: First, one could assume that a well curated and integrated IPY dataset is regarded *the* information legacy, outweighing text and (early) interpretations by far. Secondly, one could take for granted, that information within texts (e.g., journal articles and books) will be handled properly by the existing publication channels – as opposed to channels and methods for data. This second interpretation can be seen as supported by some remarks favouring open access to articles and, more importantly, by the Data and Information Management chapter insisting on "consistent and accurate acknowledgement of data sources by all data users".

## 3 eXpedition, an institutional portal to ESS

The eXpedition portal of the Alfred Wegener Institute (AWI) [6] was created as a practical means to access the output of an important class of AWI's research activities, especially that of the expeditions of the ice-breaking research vessel Polarstern. Here, eXpedition is presented as a prototype of a feature, or viewing angle, which needs to be implemented by future, comprehensive systems, namely the virtual observatories [7] for the global, interdisciplinary Earth System Science.

The special feature of eXpedition is to establish a relationship between information items – journal articles, datasets, a map, distinguished persons and some "grey" texts, e.g. press releases and weekly expedition reports - based on the expedition specified. An expedition - like some other types of *events* - establishes (or at least: makes possible) a meaningful concurrency, co-location and (possibly) co-operation of scientific disciplines, research themes, projects and people involved.


eXPEDITION Portal : POLARSTERN data - Mozilla Firefox
Datei
Bearbeiten
Ansicht
Chronik
Lesezeichen
Extras
Hilfe
http://web.awi.de/php/ResearchPlatform/Display.php?year=
Google
eXPEDITION Portal : POLARSTERN...
Publishing Network for Geoscientific & ...

| Expedition | Date \|\| Port | Region \|\| Research | Weekly reports | Data \|\| AWI Publications | Details |
|---|---|---|---|---|---|
| **ANT-XXI/3**<br>Coordinator:<br>Pörtner, H.<br>Chief scientist:<br>Smetacek, V. | 21.01.2004 - 25.03.2004<br>Capetown - Capetown<br>[Map(png)] | Atlantic/Indian Ocean, Polar frontal zone<br>Biology, EIFEX<br>[Press release] | 1. Report<br>2. Report<br>3. Report<br>4. Report<br>5. Report<br>6. Report<br>7. Report<br>8. Report<br>9. Report | Meteorology<br>PANGAEA: Stations<br>PANGAEA: Datasets<br>*[Note: Datasets for recent cruises may not yet be available]*<br>**"Polarstern Abstracts"**<br>Assmy *et al* [2006]a<br>Assmy *et al* [2006]b<br>Assmy *et al* [2006]c<br>Assmy [2006]d<br>Bathmann [2004]<br>Bathmann [2005]a<br>Bathmann [2005]b<br>Bathmann [2005]c<br>Bathmann [2005]d<br>Bathmann *et al* [2006]<br>Croot *et al* [2007]<br>Harbou *et al* [2004]<br>Harbou *et al* [2006]<br>Henjes *et al* [2004]<br>Hibbert *et al* [2006]<br>Jacquet *et al* [2006]<br>Kruse [2005]<br>Kruse [2006]<br>Leach *et al* [2004]<br>Sachs *et al* [2005]<br>Smetacek *et al* [2005]<br>Strass *et al* [2005] | ► |

Figure 1: The eXpedition portal, a prototype network of data and publications

For example, the first expedition of Polarstern in 2004, ANT-XXI/3 (see fig.1), was host to EIFEX, the second iron fertilization experiment on board Polarstern. Physical and biological oceanographers as well as chemists gathered data for a quite recent discipline or branch of science, biogeochemistry. This can be gathered most easily from the cruise reports. In particular, the first report [8] reveals that the scientific crew came from "14 institutes and 3 companies from 7 European countries and South Africa". A researcher considering a project of similar scientific complexity might find the reference to the coordinator and chief scientist of ANT-XXI/3 helpful, in order to gather advice for her own expedition. Thus, these "data", probably held in low academic esteem, can nevertheless provide valuable information as to the practical conduct of research and also for the outreach to teachers and students, especially in their role as prospective polar researchers - a function of outreach taken quite seriously, e.g., by IPY [2].

In 2005, the official report on the expeditions ANT-XXI/3-4-5 was published [9] (Unfortunately, it is not yet available in digital form, online). Mostly after 2005, many journal articles and datasets originating from ANT-XXI/3 were published. Therefore, these publications could not have been gathered by a bibliography within the timely expedition report. However, many of them – especially those authored by AWI members – can be accessed through eXpedition, thus closing a gap in the aggregation of information about Polarstern, ANT-XXI/3 and EIFEX, to some degree.

As an *institutional* portal, the main added value of eXpedition is to aggregate results of *AWI* expeditions, mostly gained by *AWI*-members, thus, e.g., contributing to the justification of expenses for Polarstern and other infrastructures, maintained by AWI. This de-facto restriction of eXpedition to the context of AWI's members' publications and research platforms limits its usefulness as a tool for science, since it is lacking completeness. The concluding chapter will discuss potential solutions to achieve an almost complete coverage of each activity in ESS, by aggregation and evaluation systems conceptually similar to eXpedition.

eXpedition's inherent limitations are due to the historically grown methods to acquire text, data and metadata, in a number of repositories developed and operated at the AWI. Links from eXpedition point to

- Homepages of researchers, generated automatically from the institutional directory, which contains contact as well as affiliation information, encoded in the eduPerson schema used by, e.g. Shibboleth [10], an authentication and authorization framework, quickly gaining the status of de-facto standard. The homepages of AWI have been mapped into documents with pertaining Dublin Core metadata and exposed via OIA-PMH [11].
- A map, automatically produced, during(!) each expedition, from 3-hourly meteorological observations reports, subsequently stored in one of AWI's web servers for dynamic content.
- Press release and weekly reports, stored in the content management system of AWI's main web server, www.awi.de.
- Publications from the institutional (publications) repository ePIC [12], which encodes metadata primarily using the Dublin Core

schema. A number of “proprietary” metadata fields have been added – among them a field designating the research platform (Polarstern) and cruise identifier. During the last years, the former publications “list” (database) has been extended to holding full texts, and exposing metadata through OIA-PMH, thus forming a standards-based repository.

- Datasets within the data publications system PANGAEA [13], which holds vast amounts of original datasets. It is the platform of the World Data Centre for Marine Environmental Data (WDC-Mare) – therefore it holds many datasets not being created by AWI-members or not related to AWI infrastructure. WDC-Mare and the other WDCs operate under the auspices of the International Council of Sciences (ICSU) [14]. In order to enable persistent references, it publishes all datasets using DOIs. Metadata within PANGAEA are encoded in a historically grown schema which is being exposed in a number of formats and protocols, especially using a profile of ISO 19115, a schema for geocoded information supported by the EC initiative for an Infrastructure for Spatial Information in Europe (INSPIRE). Expedition names are encoded as values of a field “event”, type “campaign” (a term which is frequently used for land expeditions). Recently, the exposure through Dublin Core and OAI-PMH has been added as an interface. [15]

Since three of the major types of information items from AWI’s directory and repositories have been exposed through OAI-PMH, and made known to the Scientific Commons harvester [16], all three types of objects can be found there through one query. Unfortunately, the query parsing does not seem to allow searching for an exact phrase, e.g. “ANT-XXI/3”. A similar test with Google resulted in only 18 hits.

Currently, due to its limited data sources, eXpedition is an almost purely institutional system. The short test with the Scientific Commons and Google on the other hand, shows dramatically, that any search in the “open web” for the results of a specific expedition is bound to fail, today. A more targeted search would become possible, if “tags” for expeditions would be included in metadata or full text as a unique string or name-value combination of an attribute.

As an intermediate conclusion, it is save to say that eXpedition has revealed the value of aggregating different types of information in the context of a specific, but common type of “event” of Earth System Sciences, namely expeditions. One can assume that this value would pay off as well in any case of organized international, multidisciplinary research for events like “projects”.

## 4 Future developments

The virtual observatory paradigm [7] – in the case of ESS based on latitude, longitude *and time*, as well as a staggering multitude of observables - is but one of a number of metaphors or coordinate systems by which navigation of the oceans of data and knowledge will become more manageable for human beings. Assuming that an observatory of this kind could be used to visualize all texts and datasets in the ESS domain, a number of powerful, yet exact filters would be necessary to cut down the number of observations and simulations available from a query covering any significant part of the world. Regarding exactness, more unique identifiers on well defined and meaningful categories of research *and the research process* would definitely be needed to achieve this.

Specific tools might discover trends in research, based on usage [17]. Yet other ranking and rating tools may rely on citation (graph) analysis [18], and reveal personal or the authority of publications [19,20]. These tools, and others, which try to discover communities (or "schools"), or "related information", are based – in most cases –on the authors' identities or on the relationship between texts (e.g., articles).

Recently, the growing importance of these analyses and the ambiguity of authors' names, as given in references and even in the full text of publications, has led to an emerging consensus about the need for a persistent and unique identifier for scientific authors [21,22]. We add that, looking a bit more into the future, and considering the research and publication process as a whole as well as the emerging need for fine-grained access control, especially regarding datasets, one should add the need for identifiers for groups, projects or co-laboratories.

In conclusion, we propose that some more "tags" or identifiers need to be considered. They would be of high significance and useful for aggregation and analysis of published text and data. IDs for texts and datasets themselves are quite well known in the repository community meanwhile. Using unique identifiers for persons, institutions, virtual organizations, programs and projects (or, equivalently, in most cases, grants) – and, of course, identifiers for expeditions, in the case of ESS- in addition would most probably be of high advantage.

The "natural" relationship - namely: "is result of" - between texts and scientific datasets on the one hand and projects (and expeditions) on the other is depicted in fig. 2. It shows all entities mentioned in this article, and some of their relationships.

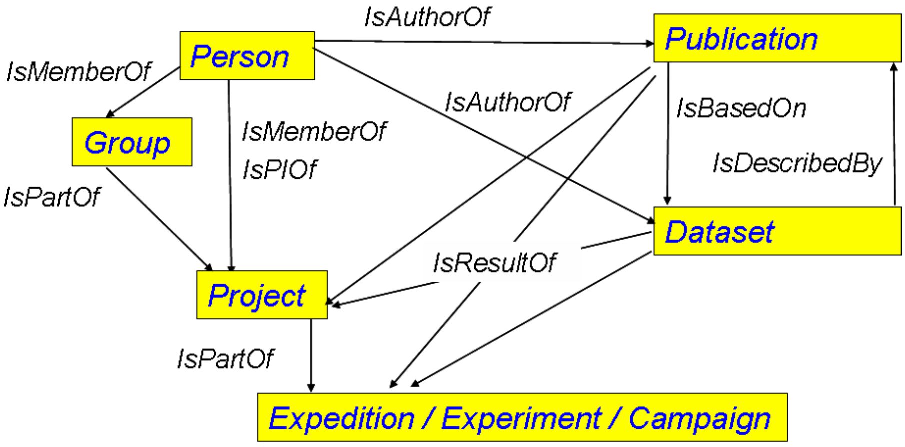


Figure 2: Tentative diagram of object relationships [21]

It remains to be determined, where – in specific sections of "full text" or in metadata fields to be specified – those additional identifiers would have to be placed. The answer will depend on a thorough investigation of text- and data-mining scenarios and methods, so that the specific assertions, encoded in the new identifiers, can be used most effectively and efficiently.

## References

1. "A Framework for the International Polar Year 2007-2008",produced by the ICSU IPY 2007-2008 Planning Group, (2004), p.10; [http://www.ipy.org/images/uploads/framework.pdf]
2. "The scope of science for the International Polar Year 2007-2008", Produced by the ICSU/WMO Joint Committee for IPY 2007–2008, (2007), p.1; [http://www.ipy.org/images/uploads/LR*PolarBrochureScientific_IN.pdf]
3. Ibid, p.13
4. IPY Website, [http://www.ipy.org/index.php?/ipy/audience/C28/] last visited Feb. 2007
5. "International Polar Year 2007-2008 Data Policy", May 2006 [http://classic.ipy.org/Subcommittees/final_ipy_data_policy.pdf]
6. eXpedition Portal [http://web.awi.de/php/ResearchPlatform/Display.php?year=2004&name=polarstern&type=ship]
7. A. Szalay, J. Gray (2001) "The World-Wide Telescope", Science (**293**) 2037, [DOI: 10.1126/science.293.5537.2037]
8. Weekly report no. 1, EIFEX, ANT XXI/3, RV "Polarstern" 25th January 2004 [http://web.awi.de/Polar/Polarstern/ANT-XXI-3/Newsletters/report1-040125-d.html#englisch]
9. V. Smetacek, U. Bathmann, E. Helmke (2005). The Expeditions ANTARKTIS XXI/3-4-5 of the Research Vessel Polarstern in 2004 (Die Expeditionen ANTARKTIS XXI/3-4-5 des Forschungsschiffes Polarstern 2004), Reports on Polar and Marine Research, Alfred Wegener Institute for Polar and Marine Research, Bremerhaven, 500, 302 pp
10. Shibboleth, an inter-institutional middleware for single-sign-on and support of authorization decisions [http://shibboleth.internet2.edu/]
11. OAI-PMH, the Open Archives Initiative Protocol for Metadata Harvesting [http://www.openarchives.org/pmh/]
12. ePIC, the institutional repository of AWI [http://epic.awi.de]

13. PANGAEA, Publishing Network for Geoscientific & Environmental Data [http://www.pangaea.de/]
14. World Data Center System [http://www.ngdc.noaa.gov/wdc/wdcmain.html] (last visited March 2007)
15. J. Brase, U. Schindler (2006) “The publication of scientific data by World Data Centers and the National Library of Science and Technology in Germany”, Data Science Journal, **5** , 205-208
16. Scientific Commons [http://www.scientificcommons.org/] last visited March 2007) ]
17. J. Bollen, R. Luce, S. Vemulapalli, and W. Xu (2003) “Usage analysis for the identification of research trends in digital libraries”; D-Lib Magazine, **9**(5)
18. Citeseer [http://citeseer.ist.psu.edu/citeseer.html] (last visited March 2007)
19. J.M. Kleinberg (1999). "Hubs, authorities, and communities." ACM Computing Surveys (CSUR), **31**(4)
20. J.M. Kleinberg, (1999) “Authoritative sources in a hyperlinked environment”, Journal of the ACM, **36**(5), 604-632
21. H. Pfeiffenberger, A. Macario (2005). „Text, Data and People – How to Represent Earth System Science”, CERN workshop on Innovations in Scholarly Communication (OAI4), 20.Oct.2005, Geneva, Switzerland [http://epic.awi.de/Publications/Pfe2005c.pdf]
22. Knowledge Exchange, Institutional Repositories Workshop Outcomes (on Author ID) (2007) [http://knowledge-exchange.net.dynamicweb.dk/Default.aspx?ID=102] (last visited March 2007)