Webservices Infrastructure for the Registration of Scientific Primary Data

Uwe Schindler[1], Jan Brase[2], and Michael Diepenbroek[1]

[1] World Data Center for Marine Environmental Sciences (WDC-MARE),
MARUM, University of Bremen, Leobener Str., 28359 Bremen, Germany
uschindler@wdc-mare.org, mdiepenbroek@wdc-mare.org
[2] Research center L3S, Expo Plaza 1, 30539 Hannover, Germany
brase@l3s.de

Abstract. Registration of scientific primary data, to make these data citable as a unique piece of work and not only a part of a publication, has always been an important issue. In the context of the project "Publication and Citation of Scientific Primary Data" funded by the German Research Foundation (DFG) the German National Library of Science and Technology (TIB) has become the first registration agency worldwide for scientific primary data. Registration has started for the field of earth science, but will be widened for other subjects in the future. This paper shall give an overview about the technical realization of this important usage field for a digital library.

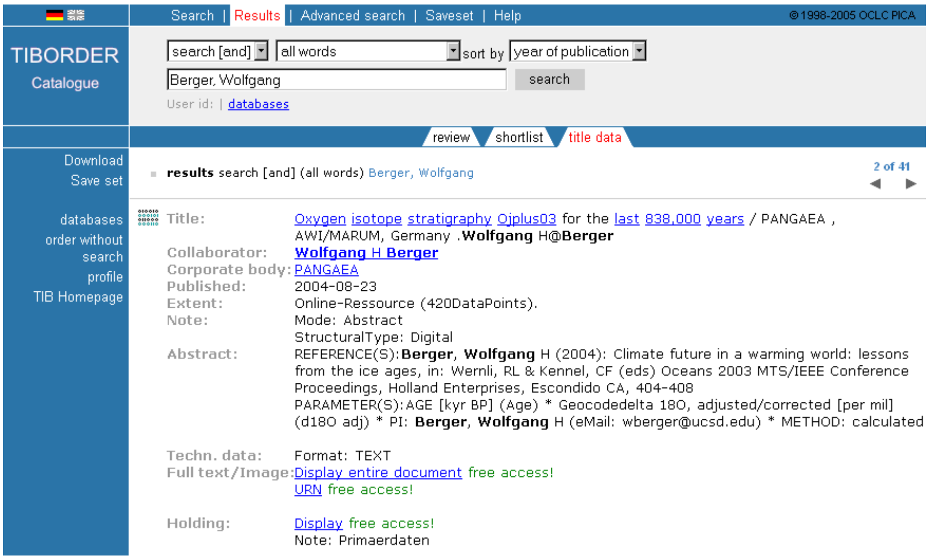
1 Motivation

In its 2004 report "Data and information", the *International Council for Science* (ICSU) [11] strongly recommended a new strategic framework for scientific data and information. On an initiative from a working group from the *Committee on Data for Science and Technology* (coData) [9], the *German Research Foundation* (DFG) [2] has started the project *Publication and Citation of Scientific Primary Data* as part of the program *Information-infrastructure of network-based scientific-cooperation and digital publication* in 2004. Starting with the field of earth science the *German National Library of Science and Technology* (TIB) is now established as a registration agency for scientific primary data as a member of the *International DOI Foundation* (IDF).

2 Registration of Scientific Data

Primary data related to geoscientific, climate and environmental research is stored locally at those institutions which are responsible for its evaluation and maintainance. In addition to the local data provision, the TIB saves the URL where the data can be accessed including all bibliographic metadata. When data are registered, the TIB provides a *Digital Object Identifier* (DOI) as a unique identifier for content objects in the digital environment. DOIs are names assigned

A. Rauber et al. (Eds.): ECDL 2005, LNCS 3652, pp. 128–138, 2005.



search [and] | all words | sort by | year of publication

Berger, Wolfgang

User id: | databases

results search [and] (all words) Berger, Wolfgang

Title: Oxygen isotope stratigraphy Ojplus03 for the last 838,000 years / PANGAEA , AWI/MARUM, Germany .**Wolfgang** H@**Berger**

Collaborator: **Wolfgang** H **Berger**

Corporate body: PANGAEA

Published: 2004-08-23

Extent: Online-Ressource (420DataPoints).

Note: Mode: Abstract
StructuralType: Digital

Abstract: REFERENCE(S):**Berger**, **Wolfgang** H (2004): Climate future in a warming world: lessons from the ice ages, in: Wernli, RL & Kennel, CF (eds) Oceans 2003 MTS/IEEE Conference Proceedings, Holland Enterprises, Escondido CA, 404-408
PARAMETER(S):AGE [kyr BP] (Age) * Geocodedelta 18O, adjusted/corrected [per mil] (d18O adj) * PI: **Berger**, **Wolfgang** H (eMail: wberger@ucsd.edu) * METHOD: calculated

Techn. data: Format: TEXT

Full text/Image: Display entire document free access!
URN free access!

Holding: Display free access!
Note: Primaerdaten

Fig. 1. A dataset as a query result in the library catalogue

to any entity for use on digital networks. They are used to provide current information, including where they (or information about them) can be found on the Internet. Information about a digital object may change over time, including where to find it, but its DOI will remain stable. For more information, we refer to [3]. In cooperation with the *German National Library* (DDB) every dataset is furthermore registered at the infrastructure of the project EPICUR [8] with a unique URN. Due to the expected large amount of datasets that need to be registered, we have decided to distinguish between citable datasets on the collection level and core datasets usually are data entities of finer granularity. Core datasets receive their identifiers, but their metadata is not included in the library catalogue whereas citable datasets, usually collections of, or publications from core datasets are included in the catalogue with metadata compatible to ISO 690-2 and Dublin Core (DC, see [10]). The DOI guarantees that these data are generally accessible and are citable inside traditional publications. By this, scientific primary data are not exclusively understood as part of a scientific publication, but have its own identity.

All information about the data is accessible through the online library catalogue of the TIB. The entries are displayed with all relevant metadata and persistent identifiers as links to access the dataset itself (see fig. 1).

3 Infrastructure

A special infrastructure is needed for flexible registration of DOIs for datasets and migration of meta information into related library catalogues. The key ele-

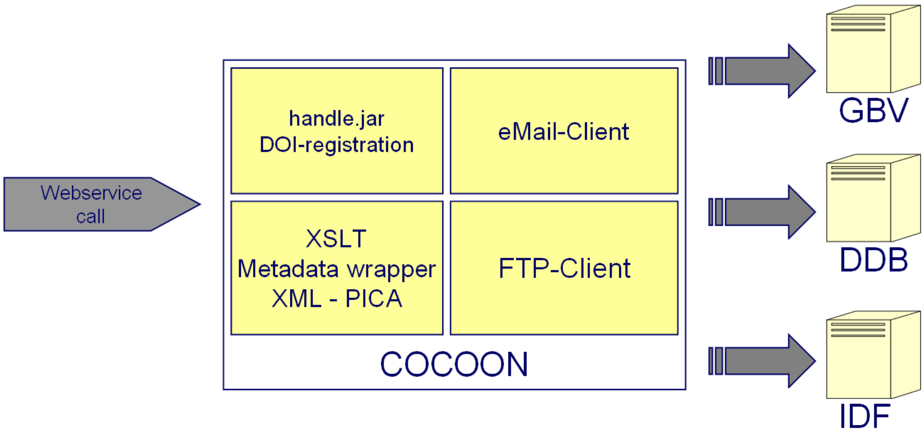


Fig. 2. The architecture of the registration process

ment is a webservice as part of the middleware at the TIB that offers automatic and manual upload of registration information.

Figure 2 gives an overview on the components and the possible workflows.

3.1 Webservice

We have choosen the SOAP (**S**imple **O**bject **A**ccess **P**rotocol) webservice standard for the communication between the data providers and the *TIB* because webservices provide interoperability between various software applications running on disparate platforms and programming languages. This is possible because open standards and protocols are used. Protocols and data formats are XML based, making it easy for developers to comprehend. By utilizing HTTP/HTTPS on the transport layer, web services can work through many common firewall security measures without requiring changes to the firewall filtering rules. This is not the case with RMI or CORBA approaches.

As the STD-DOI webservice is SOAP conformant, data providers can embed the client stub into their middleware by importing the WSDL (**W**eb**S**ervice **D**escription **L**anguage) file into their application server. For the webservice we have identified five different methods:

1. **registerCitationDOI** - For a citable dataset a DOI and an URN are registered
2. **registerDataDOI** - A core dataset only receives a DOI
3. **transformData2CitationDOI** - Upgrade a core dataset to a citable dataset by adding metadata
4. **updateCitationDOI** - If any part of metadata changes for a citable dataset, a new library record has to be created
5. **updateURL** - If the URL of a dataset changes, this information has to be stored at the DDB for the URN and the IDF for the DOI resolution

An exerpt of the WSDL-file describing the Webservice can be seen below:

```
...
<wsdl:portType name="CodataWS">
  <wsdl:operation name="registerCitationDOI" parameterOrder="xml url">
    <wsdl:input message="impl:registerCitationDOIRequest"
       name="registerCitationDOIRequest"/>
    <wsdl:output message="impl:registerCitationDOIResponse"
       name="registerCitationDOIResponse"/>
  </wsdl:operation>
  <wsdl:operation name="registerDataDOI" parameterOrder="doi url">
    <wsdl:input message="impl:registerDataDOIRequest" name="registerDataDOIRequest"/>
    <wsdl:output message="impl:registerDataDOIResponse" name="registerDataDOIResponse"/>
  </wsdl:operation>
  <wsdl:operation name="transformData2CitationDOI" parameterOrder="xml">
    <wsdl:input message="impl:transformData2CitationDOIRequest"
       name="transformData2CitationDOIRequest"/>
    <wsdl:output message="impl:transformData2CitationDOIResponse"
       name="transformData2CitationDOIResponse"/>
  </wsdl:operation>
  <wsdl:operation name="updateCitationDOI" parameterOrder="xml">
    <wsdl:input message="impl:updateCitationDOIRequest"
       name="updateCitationDOIRequest"/>
    <wsdl:output message="impl:updateCitationDOIResponse"
       name="updateCitationDOIResponse"/>
  </wsdl:operation>
  <wsdl:operation name="updateURL" parameterOrder="doi url">
    <wsdl:input message="impl:updateURLRequest" name="updateURLRequest"/>
    <wsdl:output message="impl:updateURLResponse" name="updateURLResponse"/>
  </wsdl:operation>
</wsdl:portType>
...
```

To prevent unauthorized access of the webservice we have choosen HTTPS as transport layer. In the first approach we have simple username/passwort access restrictions. In the future we want to use client authorization with certificates on the SSL/TLS layer of the HTTPS protocol. Every data provider then gets his own client certificate that can be embedded into the webservices client key store for authorization.

3.2 Metadata Scheme

We identified a set of metadata elements to describe the bibliographic information of our scientific primary data. Whenever possible, we have tried to use Dublin Core (DC) equivalent metadata elements. The metadata scheme can be found in [6] or at the project's webpage [13], it includes all information obligatory for the citing of electronic media (ISO 690-2). For inclusion into the library catalogue, however, we had to convert the XML-based metadata into PICA-format. PICA, an acronym for *Project for Integrated Catalogue Automation* is a cataloguing format based on MARC21 of the *Library of Congress*. It is used in the *Central Library Database of northern Germany* (GBV) that is responsible for the cataloguing at the *TIB*.

3.3 The Infrastructure at the Data Providers – Example PANGAEA/WDC-MARE

The DOI registration webservice client is embedded into the metadata publishing workflow of PANGAEA (Publishing Network for Geoscientific & Environmental

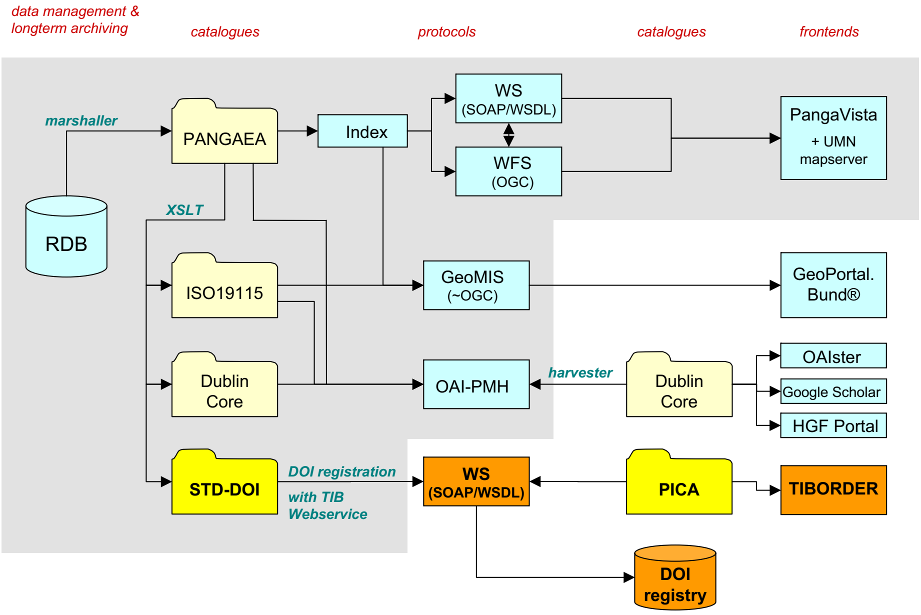


Fig. 3. The STD-DOI webservice embedded into the PANGAEA middleware structure

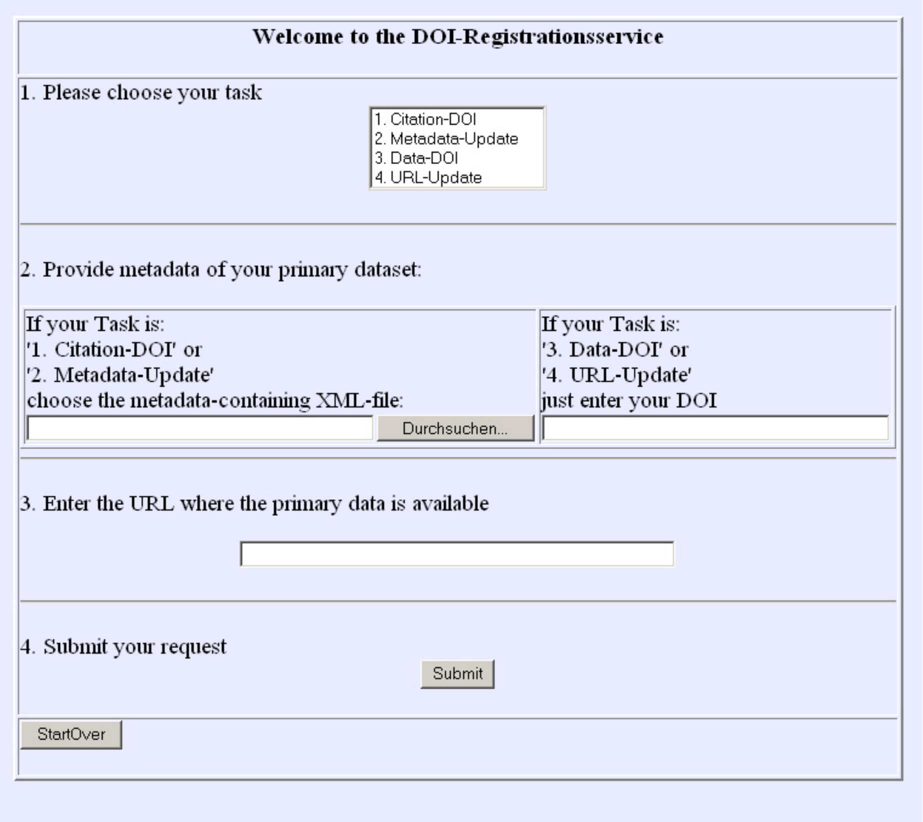
Data [17]). After inserting or updating a dataset in PANGAEA the import client queues background services which keep the XML metadata repository up to date (see fig. 3).

These background services marshal in a first step the metadata into an internal XML schema. This schema reflects the PANGAEA database structures and is optimized for simple marshalling of database records and transformation into other formats. We have choosen "jAllora" of HiT Software (see [15]) because of its rich marshalling options for this task. With this software the underlying SYBASE [16] database structure can be easily mapped to a given XML schema.

Because of the relational database structure, a change in one relational item can lead to a change in a lot of XML files. Database update triggers fill the background services queue on changes in these related tables. This keeps the "flat" XML table in synchronization with the relational data.

The internal XML is stored as binary blobs in a database table linked to the datasets. On top of this a full text search engine (SYBASE EFTS) provides fast search access to the metadata. These XML blobs can be transformed into various other schemas with XSLT [14] on the fly:

– ISO 19115
– OGC WebFeatures (for WFS)
– Dublin Core (for a OAI-PMH repository)



Welcome to the DOI-Registrationsservice

1. Please choose your task

2. Provide metadata of your primary dataset:

If your Task is:
'1. Citation-DOI' or
'2. Metadata-Update'
choose the metadata-containing XML-file:

If your Task is:
'3. Data-DOI' or
'4. URL-Update'
just enter your DOI

3. Enter the URL where the primary data is available

4. Submit your request

Fig. 4. Screenshot of the web interface to the webservice for registration of datasets

- another internal thumbnail format, which is also stored as binary blob for fast access by the PANGAEA search engine ”PangaVista” [18]
- STD-DOI for the DOI registration of citable datasets

For DOI registration another background service registers all new/updated datasets with the status ”published” after a lead time of 30 days at the TIB by the webservice described before as *core datasets*. The lead time helps preventing inadvertent registration of datasets. During this time other data curators can look after the data and metadata and make changes which resets the lead time to 30 days again.

In PANGAEA all datasets have an unique integer ID from what the DOI is created by prefixing with a static string. The URL of the data ressource is made available through the PANGAEA webserver also by the unique ID embedded into the DOI:

dataset id 80967
↓
doi:10.1594/PANGAEA.80967
↓
`http://doi.pangaea.de/10.1594/PANGAEA.80967`

After registering the core datasets the data curator can group them into a *collection dataset* (e.g. all data of a project or all data linked to a single publication) and give them a separate *citable DOI.* For that the assigned metadata gets transformed by XSLT to the STD-DOI schema from the internal XML file. Nevertheless, it is also possible to choose a single core dataset and make a citable. Due to this workflow the registering of citable PANGAEA datasets is always an upgrade of a previous core dataset (single datafile or collection) to a citable one by adding metadata at the TIB. This is done after a 30 days lead time, too.

3.4 The Infrastructure at the TIB

The webservice at the *TIB* receives XML files from the data providers, and starts the registration process:

- The DOI is registered via a java based transmission to the DOI foundation.
- For the URN registration a XML file has to be send by e-mail to the DDB.
- The metadata has to be transformed to PICA format and uploaded on a FTP server at the central library database (GBV).

To execute these different tasks, based on a single XML file, we have based the system on Apache Cocoon (see [1])

3.5 Cocoon

Cocoon is an XML publishing framework, it was founded in 1999 as an open source project under Apache Software Foundation. Cocoon offers the separation of content, style, logic and management functions in an XML content based web site (see fig. 5).

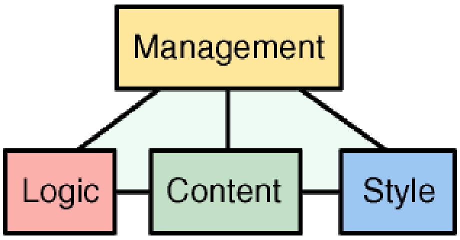


Fig. 5. Cocoon: separation of content, style, logic and management functions

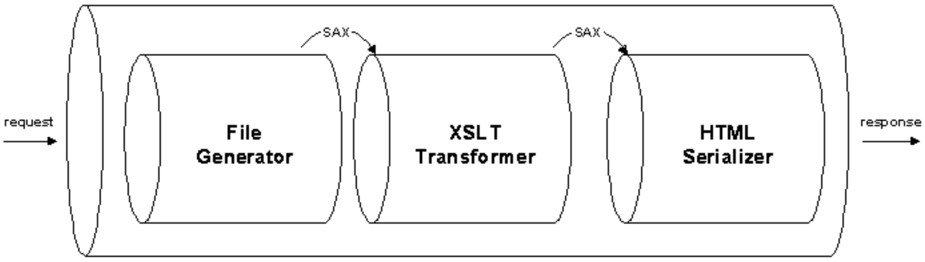


Fig. 6. Cocoon: pipeline processing

This separation allows us to easily change the parts of the architecture or the appearance of the web interface. Since it is initialised by the retrieval of a XML-file, sent to the system by the research institutes, every registration starts a XML based pipeline process (see fig. 6).

All transactions are based on XML and XSLT files.

The *eXtensible Stylesheet Language Transformation* (XSLT) is a language for transforming XML documents into other XML documents. The origins of XSL are in *Cascading Style Sheets* (CSS), where a ”stylesheet” is used to add formatting to an HTML file. The syntax to use a stylesheet in XSLT is similar to the syntax in CSS.

XSLT stylesheets have a very different function than CSS stylesheets, however. CSS allows you to define the colours, backgrounds, and font-types for an HTML web page. XSLT allows you to transform an XML file into an HTML file or another text-based format.

For a complete description of XSLT we refer to [14].

3.6 Converting XML to PICA

As you can see from fig. 2 our system also includes a translation from XML-files to the PICA format. Some example XSLT commands are:

Simple tag values. Some PICA entries can easily be derived from XML entries, or combination of XML-entries: The title in PICA (4000) is a combination of the metadata attributes title, publisher, publicationPlace and creator.

```
4000 <xsl:value-of select="/resource/title"/>/
     <xsl:value-of select="/resource/publisher"/>,
     <xsl:value-of select="/resource/publicationPlace"/>.
     <xsl:value-of select="/resource/creator"/>
```

Loops. For every author (instances of the attribute ”creator”) there is a new ordered PICA entry:

```
<xsl:for-each select="/resource/creator">
30<xsl:value-of select="position()+10"/> <xsl:value-of select="."/>
</xsl:for-each>
```

If-then structures. If the attribute relationType has the value "isCompiledBy", than the related DOI has to appear in the PICA category *4227 Compilation*, otherwise it appears in *4201 Footnote*.

```
<xsl:for-each select="/resource/relatedDOIs">
<xsl:choose>
<xsl:when test="relationType='isCompiledBy'">
4227 <xsl:value-of select="relatedDOI"/>
</xsl:when>
<xsl:otherwise>
4201 <xsl:value-of select="relationType"/>:
<xsl:value-of select="relatedDOI"/>
</xsl:otherwise>
</xsl:choose>
</xsl:for-each>
```

The XSLT code of the complete transformation can be found in the appendix of [5].

4 Status

In cooperation with

- World Data Center for Climate (WDCC) at the Max Planck Institute for Meteorology (MPIM), Hamburg
- Geoforschungszentrum Potsdam (GFZ)
- World Data Center for Marine Environmental Sciences (WDC-MARE) at the University of Bremen and the Alfred-Wegener-Institute for Polar and Marine Research (AWI), Bremerhaven

the TIB now is the world's first registration agency for primary data in the field of earth sciences.

This development will ameliorate current shortcomings in data provision and interdisciplinary use, where data sources may not be widely known and data are archived without context. It will enable citations of data in a standard manner, and also facilitate links to more specialised data schemes. The DOI system offers a proven well-developed system which is already widely deployed and enables to focus the efforts on the scientific data aspects of the project.

Authors of articles started to cite datasets using the DOI in the bibliography. One example is the following:

> Lorenz, S.J., Kasang, D., Lohmann, G. (2005):
> *Globaler Wasserkreislauf und Klimaänderungen – eine Wechselbeziehung*, In: *Warnsignal Klima: Genug Wasser für alle?* Lozán, Graßl, Hupfer, Menzel, Schönwiese (Eds.), pp. 153-158. Wissenschaftliche Auswertungen, Hamburg, Germany.

This article uses and cites:

> Stendel, M., T. Smith, E. Roeckner, U. Cubasch (2004):
> *ECHAM4_OPYC_SRES_A2: 110 years coupled A2 run 6H values*, WDCC, doi:10.1594/WDCC/EH4_OPYC_SRES_A2.

The web service installed at the *TIB* is fully functional and running. We have registered 50 citable and 200,000 core datasets so far (June 2005). All registration agents have used successfully the web interface to register datasets. WDC-MARE and WDCC are using the webservice successfully in their production environments to automatically register DOIs without any manual user interaction. We expect an amount of approximately 1,000,000 datasets to be registered by the *TIB* until the end of 2005.

We are currently discussing cooperations to extend the registration to other disciplines like medicine and chemistry. The metadata schema is flexible enough to hold entries of these disciplines. First expressions of interest came from e.g. the *European Academy of Allergology and Clinical Immunology* (EAACI, see [19]), that wishes to register the data of its content.

The possibility of citing primary data as a unique piece of work and not only a part of a publication opens new frontiers to the publication of scientific work itself and to the work of the *TIB*. The longterm availability and accessability of high-class data respectively content can be assured and may therefore significantly contribute to the success of "eScience".

References

1. The Apache Cocoon project `http://cocoon.apache.org/`
2. Deutsche Forschungsgesellschaft (German research foundation) homepage, `http://www.dfg.de/`
3. *International DOI foundation*, doi:10.1000/1, `http://www.doi.org/`
4. *The Handle System homepage*, `http://www.handle.net/`
5. J. Brase *Usage of metadata*, Ph.D. thesis, university of Hannover 2005
6. J. Brase *Using digital library techniques - Registration of scientific primary data*, In: "Research and advanced technology for digital libraries - LNCS 3232", Springer Verlag 2004, ISBN 3-540-23013-0
7. C. Plott, R. Ball *Mit Sicherheit zum Dokument - Die Identifizierung von Online-Publikationen*, In: B.I.T. journal **1** (2004) 11–20
8. *Project "Enhancement of Persistent Identifier Services - Comprehensive Method for unequivocal Resource Identification" homepage*, `http://www.persistent-identifier.de/`
9. Committee on Data for Science and Technology (coData), `http://www.codata.org/`
10. The Dublin Core Metadata Initiative (DCMI), `http://dublincore.org/`
11. International Council for Science, `http://www.icsu.org/`
12. Learning Technology Standards Comittee of the IEEE: *Draft Standard for Learning Objects Metadata IEEE P1484.12.1/D6.4* (12. June 2002), `http://ltsc.ieee.org/doc/wg12/LOM_1484_12_1_v1_Final_Draft.pdf`
13. Project webpage, `http://www.std-doi.de/`
14. W3C *XSL Transformations Version 1.0, W3C Recommendation*, `http://www.w3.org/TR/xslt`

15. HiT Software: jAllora, http://www.hitsw.com/
16. SYBASE Inc., http://www.sybase.com/
17. Diepenbroek, M; Grobe, H; Reinke, M; Schindler, U; Schlitzer, R; Sieger, R; Wefer, G (2002) *PANGAEA - an Information System for Environmental Sciences.* Computer and Geosciences, 28, 1201-1210, doi:10.1016/S0098-3004(02)00039-0
18. PangaVista search engine, http://www.pangaea.de/PangaVista
19. EAACI website, http://www.eaaci.net/