

THE CCLRC DATA PORTAL

Glen Drinkwater, Shoaib Sufi

CCLRC – Daresbury Laboratory, Daresbury, Warrington, Cheshire, WA4 4AD, UK.

E-mail: g.j.drinkwater@dl.ac.uk, s.a.sufi@dl.ac.uk

Abstract: *The project aims to provide easy, transparent access to experimental, observational, simulation and visualisation data kept on a multitude of systems and sites. Further more it will provide links to other web/grid services, which will allow the scientists to further use the selected data, e.g. via data mining, simulations or visualisation. The Data Portal will aim to work as a broker between the scientists, the facilities, the data and other services. The problem addressed is that currently the scientific data is stored distributed across a multitude of sites and systems. Scientists have only very limited support in accessing, managing and transferring their data or indeed in identifying new data resources. In a true Grid environment it is essential to ease many of these processes and the aim of the Data Portal is to help with automating many of these tasks. The Data Portal originally used Suns Java 2 Enterprise Edition (J2EE) but was re-engineered using a component based web service model as its business logic layer with JSPs as its presentation layer.*

Key words: Data Access, Globus, Grid Services, Web Services, MetaData, SRB, Data Portal.

1. Introduction

Currently the scientists are forced to manually relate between all the experimental, data, computing and analysis facilities that are available world wide, with little infrastructure support. In the future it is hoped the Grid will provide these functions, enabling the scientists to choose much more easily from a wide range of services, connecting and combining desired services for an optimal working environment. Much of the access to the Grid is envisaged to take place through customisable, community oriented Portals. A range of projects within Council for the Central Laboratory of the Research Councils' (CCLRC) have been

chosen to provide the building blocks of an integrated solution for users of experimental, computing and data facilities, showing how technologies can be used to build middleware components that support high level scientific grid applications. Data will play a pivotal role in the success of Grid or e-Science developments. Virtually all envisaged applications will need to be able to draw from and deliver to the distributed heterogeneous information/data sources with a variety of contents. Hence three major challenges are posed: data accessibility, data transfer and management of personal data. Data accessibility implies the capability to locate information/data without prior knowledge of its physical location or the form in which its contents is described. Furthermore scientists, as well as applications, need to be able to combine results from different sources. Data transfer relates to the problem of large data volumes that need to be transferred across the Internet. Management of personal data is concerned with the growing distribution of data produced by scientists within a Grid environment, which required new ways of keeping track and moving data for single scientists and more importantly for research groups. CCLRC's integrated data system includes the following components a Data Portal for high-level access to multidisciplinary data, linking to existing data catalogue systems. These catalogues include metadata as well as links to the data itself. The data itself is held in various storage resources from local disks, databases, other data resources to multi terabyte tertiary tape systems.

1.1. Current Status

The Data Portal is currently being used in two projects involving CCLRC, the e-Materials and e-Minerals [1] projects, and generic CCLRC instance giving access to data from two of our experimental departments: Synchrotron Radiation (SR at Daresbury) and Neutron Spallation (ISIS at Rutherford (RAL)) as well as data from the British Atmospheric Data Centre (BADC at RAL) and an outside source at the Max Planck Institute for Meteorology in Hamburg, Germany. The previous two are both World Data Centres. The installation is available at <http://dataportal.dl.ac.uk:8080>.

The Data Portal was designed to work in distributed and heterogeneous environments, our current installations have proven that to be true, integrating a multitude of systems types, operating systems, data resources (databases, Storage Resource Brokers (SRB) [2]) and sites seamlessly.

1.2. Architecture

The current version of the Data Portal uses a modular web services model. This is achieved using Apache's Axis implementation of the SOAP (Simple Object Access Protocol) submission to W3C. SOAP is a lightweight protocol for exchange of information in a decentralised, distributed environment. It is a XML based protocol, which defines a framework for representing remote procedure calls and responses.

Using SOAP and web services the Data Portal was decentralised into modules that represent an area of functionality. For example, the Session Manager controls users state, Authorisation communicates with the MyProxy server [3] to authorise the user to the Data Portal and Query & Reply sends queries to multiple XML Wrappers at each facility. These services were platform and language independent allowing other services (other portals or clients) to communicate with

the Data Portal regardless of the language that they were written.

Vital to this version of the Data Portal is the Lookup module. This is used for the publishing and finding Data Portal web service modules. Essentially this acts as an interface to a Universal Description, Discovery and Integration (UDDI) registry. A module would query the UDDI and receive a Web Services Definition Language (WSDL) file address for the module. This is standard to describe the technical invocation syntax of a web service. A module would use this file to invoke the web service that it needs.

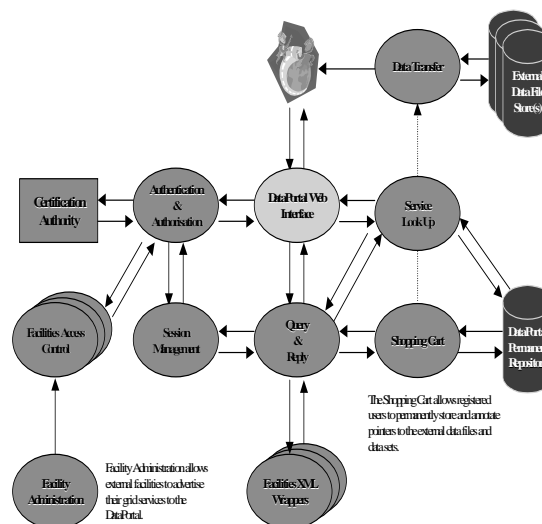


Figure 1: Web Service Architecture

The Data Portal server hosts most of these services but also provides the user interface and manages the interaction with the user and all attached resources. The server provides the user with a web interface to search the existing metadata both on the server itself and the connected data holdings transparently. Incoming requests from the user will be interpreted by the server and a query will be formed and transmitted to the facility's local repositories that are available to the Data Portal. The queries to local

repositories are XQueries [4] and are transmitted via web services. The result from the various local repositories is expected in XML format or the format that the XQuery requested (i.e. HTML). The Data Portal will collate the results and generate the required pages to display the results. The Server is also responsible for the user authentication and session control. In the future the server is also expected to liaise with other data portals as well as other grid services.

CCLRC has developed a special multidisciplinary metadata format in XML [5] to be able to integrate and make available data from various scientific topics ranging from astronomy to physics. In the following paragraph we will describe how these resources are connected with the server.

Other repositories are expected to have either their own metadata catalogue systems or their own metadata formats describing their data holding or use an extension of the CCLRC Scientific Metadata format (CSMD). To integrate them each catalogue (facility) will be accompanied by an XML Wrapper, which firstly converts the local metadata catalogue systems into CSMD. The XQuery request from the Data Portal server is executed against the CSMD and results returned to the Data Portal. Currently the metadata catalogues include links to the data location, therefore the data can be transferred via GridFTP to the user or to a third party machine.

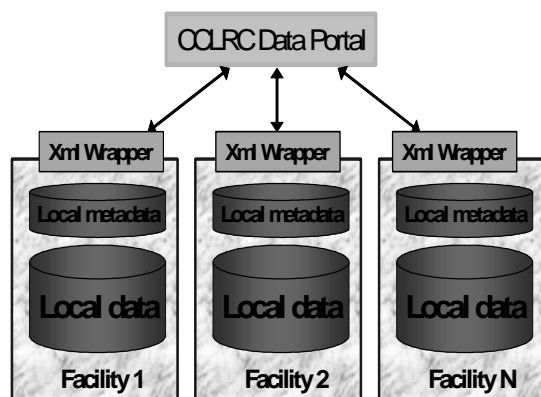


Figure 2: Simple Data Portal Architecture

The core Data Portal system offers the user the possibility to collect all relevant datasets / data files in his personal shopping basket, which can be kept from one session to the next if required. This shopping basket then offers the user a range of functionalities like transfer (using GridFTP, download), delete (from shopping basket), or if available offer other grid services to the type of data.

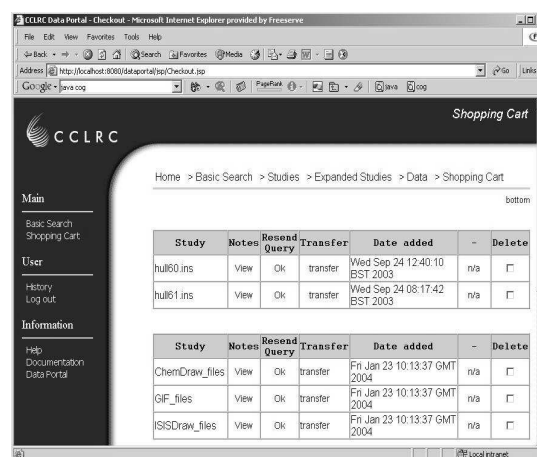


Figure 3: Shopping Basket screen shot

1.2.1. Authentication

Authentication of user to the Data Portal is via a MyProxy server. MyProxy is an online credential repository for the Grid. Storing Grid credentials in a MyProxy server allows retrieval of proxy credentials whenever and wherever needed, without worrying about managing private key and certificate files. A user would store a delegated set of credentials into the MyProxy using the Java Cogkit [6] and protects the private key with a passphrase in which is used to encrypt it. The passphrase and certificate-based authentication is then required to retrieve credentials from MyProxy, in our case the Data Portal retrieves the proxy credential for the user to access Grid resources on the user's behalf.

The Data Portal uses a MyProxy GUI Tool built using the Java Cogkit to allow

users to upload their credentials to the MyProxy for up to 30 days. This application allows scientists to easily upload their credential without installing Globus or any other software, giving them a simple GUI to create, upload and look after their credentials.

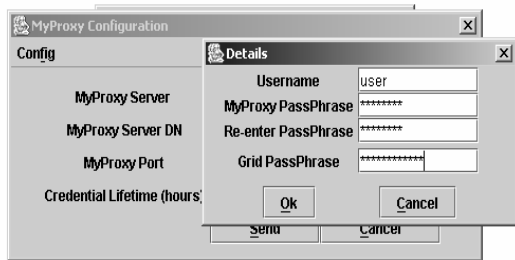


Figure 4: Data Portal MyProxy Tool

This application is launched using Java Web Start, a technology allowing a user to deploy a standalone Java software application with a single click over the network. The user would only use this application on their own secure machine with their credentials locally stored. Once their credentials have been uploaded via GSI security to the MyProxy, they can access the Data Portal from any other machine connected to the internet by passing their username and passphrase to the Data Portal which in turn retrieves the proxy credential for the user from the MyProxy. The user would then be authenticated to the Data Portal.

1.2.1.1. Single Sign On with other services

One of the key reasons for moving to web services is for the ability for other applications and web services to access the Data Portal's services. This is independent of the programming language that the web services are written in. At CCLRC two other portals have been developed, the HPC portal and Visualisation portal. The HPC Portal is to search for resources and submit HPC applications to a computational Grid and the Visualisation Portal to view data and results.

One possible scenario is that a user has found some data using the Data Portal and wishes to run an application on the data and visualise the results. The web service infrastructure of the Data Portal allows other applications to communicate with the modules. The HPC portal could communicate with the Data Portal's Shopping Basket on behalf of the user. Once the location of the data was extracted from the basket, the data could be transferred to a machine where the application resides. Once the application finishes the results can then be visualised via the Visualisation portal which can be invoked again using web services.

Single sign on between the HPC Portal and the Data Portal is accomplished by session information being shared between the separate Session Managers of the portals. A user wishing to use the functionalities available to the Data Portal e.g. a file URL in their shopping basket, who is coming through the HPC Portal, could use their proxy credential stored in the HPC Session Manager to authenticate them to the Data Portal with GSI delegation of the user's credential. This is achieved first via mutual authentication between the two web servers. Once the Data Portal Session Manager has authenticated that the client is the HPC Portal, it trusts the delegation of the credential to the Data Portal Session Manager and starts a session on the Data Portal.

1.2.2. Authorisation

The authentication of data within the Data Portal [7] is done by the GSI delegation of the user's proxy certificate to each facility. At each facility sits an Access & Control (ACM) and a XML Wrapper web service. Upon logging on, the user's proxy certificate is delegated to each facility's ACM. The ACM maps the user's distinguished name (DN) to a local user on their system and the access rights are given back to the Data Portal in the form of an XML document. The XML document gives information regarding the

read access to the facility, data, metadata respectively and other information like the user's DN, lifetime the user's access rights, normally 2 hours to match the proxy certificates lifetime. This XML document is known as an Authorisation Token.

```
<?xml version="1.0" encoding="UTF-8"?>
<attributeCertificate>
  <aclInfo>
    <version>1.0</version>
    <holder>/C=UK/O=eScience/OU=CLRC/L=DL/CN=gle
n drinkwater</holder>
    <issuer>EMAILADDRESS=ca-operator@grid-
support.ac.uk, CN=CA, OU=Authority, O=eScience,
C=UK</issuer>
    <issuerName>ACMEMIN</issuerName>
    <issuerSerialNumber>1</issuerSerialNumber>
    <signatureAlgorithm>SHA1withRSA</signatureAlgorithm>
    <validity>
      <notBefore>2004 0 27 13 35 28</notBefore>
      <notAfter>2004 0 27 14 38 10</notAfter>
    </validity>
    <attributes>
      <DPView>t</DPView>
      <wrapperGroup>t</wrapperGroup>
      <dataAccessGroup>t</dataAccessGroup>
    </attributes>
  </aclInfo>
  <signature>e/CUhswwg6yhnI9/+gGbiTB9o0dcsijlE19
PmODbjjPB3JntF4+3OMkB+uKliwXd5xVGa9AEH/HrHca+
3/qiRJPu</signature>
</attributeCertificate>
```

Figure 5: Format of Authorisation Token

The ACM signs the XML document with the facility's private key and sends the Authorisation Token via SOAP back to the Data Portal which stores it in a database. When the Data Portal sends a query to the XML Wrapper, it also sends the Authorisation Token. The XML Wrapper can validate the signature of the Authorisation Token with the facility's public key. Therefore the XML Wrapper can trust the access information regarding the facility given in the Authorisation Token.

The access information held within the Authorisation Token is specific to the facility that created that token and is only passed back to the XML Wrapper corresponding to the facility. Once trust is gained, the access information can then be used for the user. This architecture allows

each facility to have fine grained authorisation of users to large database group role authorisation. For example, one facility may wish to have two users on there meta-database, read and no access. Either the user could access the metadata in the database or not. How the facility defines the groups in the Authorisation Token is down to the facility as only the ACM and XML Wrapper will see or even trust this information. On the other hand, the facility may wish to assign a fine grained role approach for each user, each with different access rights (or the case for no access role).

Initial log on through Browser / Outside Service

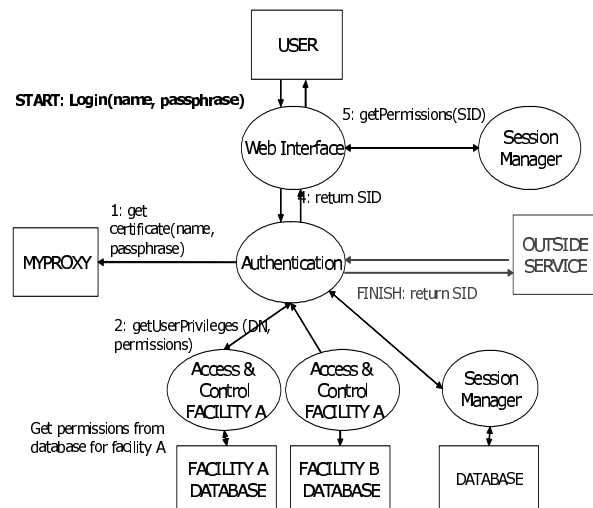


Figure 6: Logging on Process.

1.2.3. CCLRC Scientific Metadata format

A Common model for the representation of scientific study metadata does not exist, by proposing a model and an implementation, the adoption of such a system would aid interoperability of scientific information systems on the Grid, or at the very least the model will form a specification of the type and categories of metadata that studies should capture about their investigations and the data they produce. This

will allow further exploitation of the Study, associated datasets, ease citation, facilitate collaboration and allow the easy integration of pre-Grid metadata into a common Grid based information platform.

The CCLRC scientific metadata model is study-dataset orientated model and comprises of information pertaining to provenance, conditions of use, data description and location and related material, and includes indexing information. The main influences for developing the model were in-house facilities at CCLRC; specifically ISIS (Neutron Spallation at Rutherford Appleton Laboratory), SR Synchrotron Radiation source (at Daresbury Laboratory) and the British Atmospheric Database (BADC) at RAL.

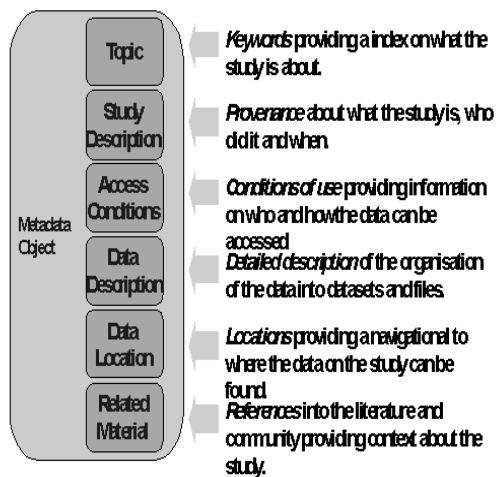


Figure 7: Schematic of metadata model

The specific metadata formats which have influenced the design and ordering of the CSMD are CIP from Earth observation [8], DDI from social sciences [9], publication type metadata from the Dublin Core [10] and lower level 'Scientific Data Objects' metadata found in XSIL [11] as well as CERA [12] from the MPIM in Hamburg. The Dublin core was found to be too high level and not detailed enough whereas XSIL lower level and missed higher level entities, CERA

was a close fit but was somewhat specific to the Earth Sciences and as generality was a key feature of our metadata model, CSMD was developed.

The Model Specifies in a semi-structured way the types of metadata that need to be captured which will make studies easier to exploit, cite, groups to collaborate and allow a lowest common denominator for scientific study Information integration within a Grid environment.

1.2.4. XML Wrapper

The XML Wrappers are used to convert between the local metadata format of a data archive and the CSMD. This allows the integration of metadata repositories which hold information about scientific studies and their associated data in various different formats and present a common interface to them via web services, which allows the Data Portal to seamlessly interact with the different data archives.

It allows the composition of queries to different data archives by placing a wrapper (or adaptor) around the archive which translates the native structure of the data into a form that the Data Portal core modules can understand. It converts the data into a common xml representation and then allows either bringing back the whole data representation or by applying a modified XQuery to bring back relevant data allowing easier and uniform composebility by providing a common interface; once the data format of the Data Portal is known (the CSMD) then XQueries can be written against this to extract relevant metadata. Thus web interfaces can be built using this common set of XQueries allowing for extensive web portal functionality.

The XML Wrapper has two independent aspects: the building of XML documents into CSMD documents and querying those XML documents.

1. **XML Wrapper Document Builder:** Selects data from the Data Archive and builds CSMD records and inserts the validated records into the XML document Repository.
2. **XML Wrapper Document Selector:** processes incoming queries on the built XML documents to retrieve relevant scientific metadata.

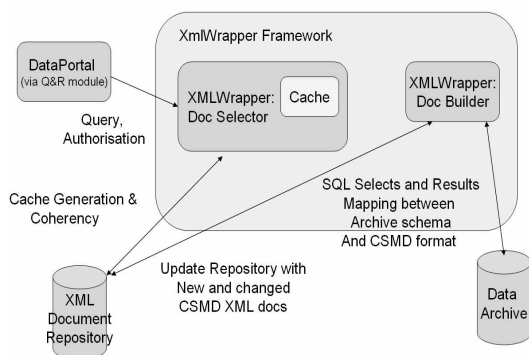


Figure 8: XML Wrapper Architecture

1.2.4.1. Document Builder

The document builder maintains a connection to the archive and periodically selects data from the archive and builds up record of the data into the Meta Data Format, which is loaded into a XML Document Repository. For Cache Coherency, it monitors changed records in the data archive and adds, updates (and perhaps deletes) documents in the repository to be consistent with the metadata held in the data archive

1.2.4.2. Document Selector

This is exposed as a web service to the Data Portal. It takes four arguments, two for the authorisation and two for the query. The first two are for the proxy credential and the Authentication Token described before. The Document Selector checks that the

Authorisation Token is signed by the facility's private key, if so, it checks to see if the DN from the certificate is the same as the DN from the Authorisation Token. If all the security steps are completed the Document Selector obtains the user's role from the token within the facility's system and is then able to query the XML document repository for the results from the user's role privileges.

The next two arguments are for the XQueries the user wishes to apply for their search. The first executes an XQuery on the XML Document Repository to obtain the results, the second is a formatting option that can change the results into HTML or different XML formats, like an XSLT transformation. After the transformation the formatted results is sent back to the Data Portal to be displayed to the user.

The benefits of this type of architecture are that the wrapper allows seamless access to the archive even if the data archive at the facility is down and unavailable. The Document Selector can still process queries and retrieve results using data from the XML Document Repository.

1.2.5. Data Storage

The Data Portal is able to download or transfer data from a variety of storage systems. Currently the Data Portal can use GSI delegation to access data stored in FTP, HTTP, GASS and GridFTP servers. But recently the Data Portal was able to access data stored in a Storage Resource Broker (SRB).

The SRB is a software product developed by the San Diego Supercomputing Centre (SDSC). It allows users to access files and database objects seamlessly across a distributed environment. In simple terms, the actual physical location and way the data is stored is abstracted from the user, who is presented with an interface similar to a regular file system.

Access to the data can be through GUIs, web browsers but also web services or

SRB APIs. This allows the Data Portal to access data sets or files through web services and download them for the user. The access and control is done by SRB. When the data is put into the system, the user specifies which users on the SRB system can have access to the data, like a Windows File System, allowing read, write, delete etc access for each file or database blob.

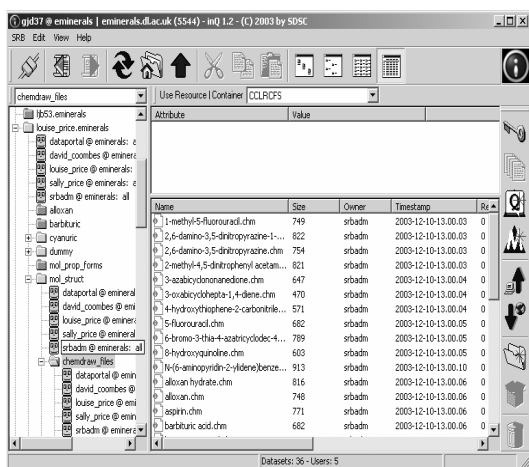


Figure 9: InQ showing a SRB view

2. Future

The Data Portal is will be taking account of new technologies (e.g. Java Portlet API, Java Server Faces, WS-Notification, WS-Resource Framework and Globus Toolkit 4). Further work and research will be undertaken with other projects and the new technologies mentioned above. This will allow new additional web service modules as well as improvements to the current modules.

2.1. GSI SRB integration

SRB version 3 has the ability to allow access using GSI authentication to the SRB through web services. The Data Portal delegate the user's proxies certificate to authentication themselves to SRB for file upload or download/transfer.

2.2. Advanced Searching

Using XQuery as the language to search through the XML Documents allows detailed advanced searched to be added to the Data Portal search capabilities. The amount of detail within the CSMD representing the scientific metadata is the only limiting factor in how specific the advanced search can be. This allows the ability for the user to specify their own XQuery or through a GUI to tweak a pre-made advanced search for certain dates, PIs, Institutions etc.

2.3. Shopping Basket sharing

Users will be able to give 'tickets' to other users of the Data Portal allowing restricted access to their Shopping Basket and the information held. This would be useful for PIs giving access to their basket to fellow researchers or post graduates.

3 References

- [1] Blanshard L, Environment from the Molecular Level e-Science project and its use of CCLRC's Web Services based Data Portal. http://www.e-science.clrc.ac.uk/documents/staff/kerstin_kleese/e-min2003.doc
- [2] Dr Michael Doherty, SRB in action. http://www.e-science.clrc.ac.uk/documents/projects/storage_resource_broker/srb_in_action.pdf
- [3] MyProxy Server. <http://grid.ncsa.uiuc.edu/myproxy>
- [4] W3C, XQuery. <http://www.w3.org/XML/Query>
- [5] Matthews BM, Sufi SA. The CCLRC Scientific Metadata Model - Version 1. The CCLRC Scientific Metadata Model - Version 1 2002. <http://www-dienst.rl.ac.uk/library/2002/tr/dltr-2002001.pdf>
- [6] Java Cogkit. <http://www-unix.globus.org/cog/java>
- [7] Ananta Manandhar. Grid Authorisation Framework for the CCLRC DataPortal. <http://www.e->

- science.clrc.ac.uk/documents/staff/kerstin_kleese/
Authorisation.pdf
- [8] CIP Metadata from Earth Observation.
<http://lcweb.loc.gov/z3950/agency/profiles/cip.html>
- [9] DDI Metadata from Social Sciences.
<http://www.icpsr.umich.edu/DDI>
- [10] Dublin Core Metadata Initiative.
<http://dublincore.org/>
- [11] Extensible Scientific Interchange
Language.
<http://www.cacr.caltech.edu/SDA/xsil/>
- [12] Meta information on Georeferenced Data.
<http://www.pikpotsdam.de/dept/dc/e/sdm/cera/>