Applied Information Technology opens Virtual Platform for the Legacy of Alexander von Humboldt

Prof. Dr. rer. nat. Detlev Doherr
Fakultät Maschinenbau
und Verfahrenstechnik (M+V)
Leiter des Instituts für Wissenschaftliche Weiterbildung

Badstraße 24
77652 Offenburg
Tel.: 0781 205-281
E-Mail: doherr@hs-offenburg.de

1953: Geboren in Göttingen
1983: Promotion zum Dr. rer. nat.
1983–1990: Geologe bei Kali und Salz AG, Kassel, und Projektleiter für die Entwicklung eines Geoinformationssystems mit IBM Deutschland GmbH
Seit 1990: Professor für Informatik und Umweltinformatik an der Hochschule Offenburg
1993 – 2010: Leiter des Hochschulrechenzentrums sowie Leiter des Steinbeis-Transferzentrums „Informationssysteme“ (früher „Umweltinformatik“)
Seit 1998: Zertifizierung zum European Geologist

Forschungsgebiete: Informationssysteme und Geoinformationssysteme, digitale Bibliotheken, E-Learning für Weiterbildungseinrichtungen und Berufsverbände, nachhaltige Entwicklungen im Bereich der Geothermie, Modelle und Simulationen zu den Energiemärkten für Erdöl und Erdgas

---

4.5 Applied Information Technology opens Virtual Platform for the Legacy of Alexander von Humboldt

Prof. Dr. rer. nat. Detlev Doherr
Armand Brahaj MSc. [1]

Abstract
The Humboldt Digital Library (HDL) is a project that aims to provide digital access to the legacy of Alexander von Humboldt. The HDL runs on an open source library developed in the Hochschule Offenburg and provides a virtual research environment in which researchers can work more effectively.
This article presents the development made in the HDL to provide alternative ways of content dissemination through the OAI protocol.Through the implemtantion of the OAI-PMH data provider in the HDL, the library is accessibly in many universities and research centers everywhere around the globe.

Introduction

Since a couple of years, here at the Hochschule Offenburg we have been supporting and developing research in the area of Digital libraries through the Project Humboldt Digital Library (HDL). The HDL project aims to gather all the publications of Alexander von Humboldt and publish them in a digital format. So far we have published more than 30 volumes from Humboldt's work and more volumes are under process as we speak. The Library includes a range of texts, tables and images, as well as many tools that assist mining the data and navigating the system [3]. While designing our digital library, we have chosen to be less orthodox to the digital library concepts of presentation, but would adhere to all the non-written standards of digital libraries for a technical collaborative process of dissemination of content, including meta-data that provide specific harvester engines with information about the digital publications.

This article deals with the developments in the HDL to provide dissemination tools for the content of the HDL. The goal of the project was the implementation of the Open Archives Initiative – Protocol for Metadata Harvesting (OAI-PMH). This implementation will allow the information about the content in the library to be disseminated within the OAI harvester network.

Open Archives Initiative

In order to disseminate content from a digital library, specific data structure should be presented and made available in a web-service way. The Open Archives Initiative (OAI) [4] supports the interconnection of scholar, academic and research repositories, disseminating their contents along its network of harvesters. Implementing the Open Archives Initiative –Protocol for Metadata Harvesting (OAI-PMH) within the Humboldt Digital Library opens the gates of the library to several harvesters around the world and makes easier to find the documents that the library hosts. The OAI- Protocol for Metadata defines a mechanism for harvesting XML-formatted metadata from repositories of digital libraries, which is essential for the dissemination of the HDL project. In figure 4.5-1 you can identify the search paths for Humboldt's documents, using the services of OAI.

The OAI-PMH operates the collaboration of Service Providers and Data Providers. Service providers (harvesters) use the protocol for harvesting and storing metadata. Data providers (repositories) provide free access to the metadata describing their resources.

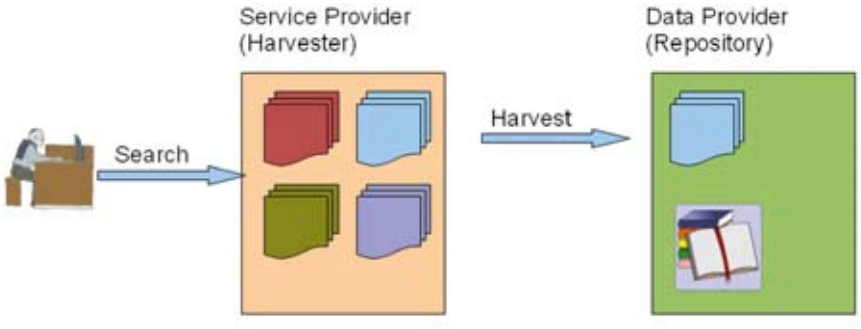

Abb. 4.5-1: *Service Provider – Data Provider model for the HDL using the OAI Protocol*

The interest of the developments in the HDL project was to provide a Data-Provider Service that could be easily harvested by many of the Service Providers out there [5]. The service should comply with all the requirements of the OAI and allow other dynamic information to be added as the system grows.

The service providers play a similar role to the role that search engines play on internet. They gather (harvest) information from data providers and index this information for later use. Visitors or other service providers can fetch the information from the service providers. Large list of service providers can be found on the web page of OAI.

Data Provider

To allow the HDL content to be indexed from service providers a separate service section should provide information about all the digital objects in the library. The information that will be published through the data provider has to be in a simple language that the service providers can understand. Luckily the OAI has detailed information on the protocol requests and responses. The system works on the concept of web-services (based on XML) that offer interaction between data provider and harvester. The information exchanged for each digital object is based on the Dublin Core [5] format to express the meta-data.

Technically the repository system should response autonomously to the requests made by the harvesters, which aim to collect the metadata and provide them to academic searches. Once the metadata is collected from the repository, it is available at the harvesters for user search. The searches on the harvesters use the data collected but do not interact directly with the repository. The requests collecting metadata are sent to the repository in a HTTP format using the POST or GET method, so the system should be able to read them and understand them. The system creates an XML-formatted response which should be sent back to the harvester.

The system (repository) should response autonomously to the requests made by the harvesters, which aim to collect the metadata and provide them to academic searches.

Once the metadata is collected from the repository, it is available at the harvesters for user search. The searches on the harvesters use the data collected but do not interact directly with the repository. The requests collecting metadata are sent to the repository in a HTTP format using the POST or GET method, so the system should be able to read them and understand them. The system must create an XML-formatted response which should be sent back to the harvester.

The tool should at least meet the „Minimal Repository Implementation“ of the protocol described by the OAI Community. The implementation must at least meet the following features:

Dublin Core format: As mentioned before, there are several formats for metadata dissemination. Repositories can implement one or more formats (Dublin Core, Marc21).

However, the Dublin Core format is the standard for the OAI Community and the metadata records must at least be available in this format.

Containers: The metadata record is formed by three containers, namely header, metadata and about. The latter, even though it can be very helpful when dealing with rights and provenance of the resources, is not required.

Sets: The grouping of resources into sets is helpful to Data Providers in order to narrow their harvesting. This feature is especially important whenever a repository has many resources or the subjects of the resources are diverse. However, implementation of Sets is not mandatory.

Response Compression: Compressed responses can be sent to harvesters. It would take less time whenever a big amount of metadata is to be delivered. However, not all harvesters are able to process these responses. This feature is also not compulsory.

Flow Control: It is specially used when a repository stores many resources. If the list of records is longer than hundred items, it can be partitioned into pieces and sent. This feature is not compulsory and is not necessarily taking into account the size of the repository.

The user / client defined for this protocol is normally a machine (harvester) that collects metedata form the OAI network repositories. These harvesters / clients include a xml parser which take the metadata out of the xml file.

The software writes an XML file from the PHP script. There is one root element within which all other elements are contained. A set of descriptive variables (relevant to the library identification) is defined at the beginning. The query string is taken either from the GET or the POST method. In the case there are no strings or the string contains invalid or repeated arguments, the routine leads to an error state. The string is split into key/values and stored in variables. Depending on the service requested (verb), the routine takes a different path. The process for every verb is quite different, as every verb has a different function within the system. All verb functions include error protection routines, in case the parameters are given wrongly.

The algorithm of the software tool is composed of a main common part (through which all requests may go) and a verb-specific part (depending on the request).

The main common part realizes tasks such as

- Write the XML headers
- Write the beginning of the OAI-PMH main element
- Defining the variables which will be used for describing the library
- Configure the access variables for the database
- Write the <responseDate> element
- Verify the presence of a data in the request
- Split the string into verb and its arguments
- Check for repeated verbs or parameters on the string
- Import string values to global variables

According to the service requested, the tool will run a specific sub-routine. The possible services are: Identify, ListMetadataFormats, ListSets, ListIdentifiers, ListRecords and GetRecord. Some functions, such as „prefix“ and „fucheck“, were written in order to verify the validity of the parameters. Finally error function takes control whenever an error occurs.

Implementation
For the implementation of this system, several technologies are used.

The server (avhumboldt.net) runs MySQL as database system. The database system contains a database which is part of the Humboldt Digital Library. This database contains several tables with different functions within the library. One of these tables is used only with the OAI-PMH implementation. This table is called "DokumenteOAI".

The columns represent the different metadata items regarding the resources in the library. The rows represent the different volumes contained in the repository. This database table can be seen as the source of information for the PHP script, as the script requests the values of the metadata to the database.
An example of a recordset used in data provider is as follows:

DOK_ID: 2601 – Local ID, it is valid only within the repository
Title: Political Essay on the Kingdom of New Spain Vol.1 – Name of the resource
Creator: Alexander von Humboldt – The writer of the document
Date: 1811 – The date when the document was written
Description: (A brief description of the document)
Contributor: John Black – A person, who has helped to write, transcribe or translate the document
Coverage: America – Describes the place which is relevant to the resource
Format: application/pdf – The document is a PDF file
Identifier: oai:avhumboldt.net:2601 – Global ID, valid within all the OAI network and in all the harvesters members of the network
Language: eng – The resource was written in English language
Publisher: LONDON: Printed for Longman, Hurst, Rees, Orme, and Brown; and H. Colburn: and W. Blackwood, and Brown and Crombie, Edinburgh. 1811 – Indicates the date when the resource was first published
Resource identifier: http://www.avhumboldt.net/avhdata/Political%20Essay%20on%20the%20Kingdom%20of%20New%20Spain/Vol1/Complete/Vol1_complete.pdf –Indicates the location of the HTML and PDF files
Subject: Travel to the Americas – Contains a keyword which can be used to search for the resource
Type: Text – In this case the resource is a text. Other possible values for this field are picture or video
Datestamp: 2010-01-03 – It is the date when the resource was fed into the system.

Abb. 4.5-2: *Printscreen from the University of Ilinois Search engine on the OAI Service Provider*

The dataset is retrieved by a scripting language (PHP in our case) and an output is generated in an appropriate format in XML.

The response file must be explicitly declared as XML-formatted. From the dataset and the operations with the datasets, a list of response xml urls is declared.

Evaluation
Since September 2010, the HDL Project has a data-provider interface that can be queried from any service provider based on OAI-PMH. The operation of the data provider tool divided in three stages:

- The harvester (service provider or client) sends a request to the server. This request includes a string specifying the type of service to provide
- According to the service requested, the tool searches the required information in the database
- Using the information from the database, the tool writes an XML-formatted file which is displayed to the harvester

The user / client defined for this protocol is normally a machine (harvester) that collects meta-data from the OAI network repositories. These harvesters / clients includes an xml parser which take the metadata out of the xml file.

The software writes an XML file from the web-scripting language script. There is one root element within which all other elements are contained. A set of descriptive variables (relevant to the library identification) is defined at the beginning. The query string is taken either from the GET or the POST method. In the case there are no strings or the string contains invalid or repeated arguments, the routine leads to an error state [2].

By the end of the last year, the HDL has been registered on more than 50 academic service providers. The registration has been completed successfully and at the moment we are noticing the first indexing of the HDL in the academic networks through the world. By being indexed from OAI harvesters, the chances of having the content published in the HDL being found and accessed increase significantly. One of the latest providers that has indexed our library is for example the University of Illinois in the USA (see figure 4.5-2). Any researcher searching for Humboldt within this University, will be pointed to the HDL project developed and maintained at Hochschule Offenburg. Some tests can be done by visiting the following URL: http://gita.grainger.uiuc.edu/registry/
This is a big step forward to provide the

legacy of Alexander von Humboldt on the Web by using the prototype of the virtual research library, which is developed at the University of Applied Sciences Offenburg.

Continuing research work is ongoing in providing alternative collaboration methods from our digital library to different harvesting engines on the net.

References

[1] PhD Candidate Humboldt-University zu Berlin

[2] Bejarano A.: Implementation of OAI-PMH protocol for Metadata Dissemination on The Humboldt Digital Library. Hochschule Offenburg 2010

[3] Doherr D., Brahaj A.: Information Management beyond Digital Libraries: Alexander von Humboldt in the Web. PIK - Praxis der Informationsverarbeitung und Kommunikation, 161 – 166, 2009

[4] Initiative, O. A.: The Open Archives Initiative. Retrieved 01 24, 2011, from The Open Archives Initiative: http://www.openarchives.org/OAI/openarchivesprotocol.htm (2002, 01 01)

[5] A list of Service Providers http://www.openarchives.org/service/listproviders/html

[6] The Dublin Core® Metadata Initiatives http://dublincore.org/