

Virtual Research Environment in the Digital Library of Alexander Humboldt

Prof. Dr. rer. nat. Detlev Doherr

Fakultät Maschinenbau
und Verfahrenstechnik (M+V)
Wissenschaftlicher Leiter
des Hochschulrechenzentrums

Badstraße 24

77652 Offenburg

Tel.: 0781 205-281

E-Mail: doherr@fh-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

Seit 1993: 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



4.4 Virtual Research Environment in the Digital Library of Alexander Humboldt

*Prof. Dr. rer. nat. Detlev Doherr
Armand Brahaj, M.Sc. [1]*

Abstract

After approximately 200 years, a comprehensive access to the texts of Humboldt's extraordinary exploration of the Americas is within sight. To open the legacy to the public for free access the Humboldt Digital Library (HDL) project has been developing a dynamic amount of data related to studies of Alexander von Humboldt. The library includes a range of texts, tables and images, as well as many tools that assist mining the data and navigating the system.

By introducing Web 2.0 technology implementations, we are providing new interactive and knowledge generation channels toward the concepts of new Virtual Research Environments.

The perception of the information architecture is changing in the era of Web 2.0. This new epoch of internet has proved that the community can generate valuable information. It is the duty of computer specialists to use this information, to extract knowledge and present this knowledge as per user's interest base.

In this article we argue that by providing more information retrieval capabilities, and by creating an academic network around our library, we can add another dimension to the information architecture of a digital library.

Information Technology Beyond a Digital Library

Transiting from a traditional library in a digital library is not a simple relation. In the traditional libraries, the information architecture is well defined, simple and linear.

The concepts of the traditional library can easily applied to the digital structure (since they are simple) by digitizing books and provide them in a document form, but the prospect of a digital environment is wider.

The search of a specific topic in a library starts with some metadata describing each document in the library (cue cards).

Beside the classical metadata describing each document, the digital library could integrate the content of all documents into an information network, which make it possible to define multivariate search paths and comparisons between different documents, no matter in which language or document type.

These implementations have been the basis of digital libraries for many years now. As time goes by, we look behind and we feel progress has been made in digital libraries, but progress is still to be made in the information retrieval process and results. With the increase of the information volume in digital libraries, there is also an increase of noise in the search for knowledge. Precious data is becoming every day more difficult to be found.

Internet algorithms like "Page Rank", HITS and similars provided great value in the search results and search techniques. As time passed by, it was clear that these algorithms where searching in the surface leaving a lot of information out of sight. The notion of "deep-web" emerged to point out that a lot of data is left out and more is to be done in the information retrieval discipline

Some promising researches and implementations are related to the deep web in the field of semantic search. The Google Search Engine is actually using these techniques.

The semantic approach is one of the variables that have to be taken in consideration when trying to solve the problem of the "deep web". We will refer to the search that exposes the data in the "deep web" as vertical search.

The choices and the combination of the factors that influence the vertical search is very high (it can considered infinite, because we are actually searching for knowledge and knowledge includes also a random chance of discovery), Therefore no system yet has fully delivered a state of art solution for this search.

The sense behind the vertical search is that visitors search for a "theme" and not for a term. By following this logic we have enhanced our digital library with a rich information retrieval (IR) module based mainly in the user interactions (user behaviours).

A Virtual Research Environment

The nucleus of every digital library is the information retrieval module. A key pro-

erty is the correctness that the library provides to the visitor for finding the required information. The visitors are either rewarded by the riches of the options provided by the IR of the digital library, or in an opposite scenario they are limited to what these IR module confines them to do.

As shown in the figure 4.4-1 the information in the Humboldt's digital library and network is transformed in four layers. The information floats from the repository, into services, AI filter and then to user interface. The same route is done when information from the users is entered into our the system.

This section makes a brief presentation about the Case-Based Reasoning (CBR) impact in the Information Retrieval.

The visitors of a digital library jump around the space of the digital library in search for the correct information. While they move through links, between documents, while they make notes about facts they found, they are leaving disconnected traces of what they read and how they interact with the system.

When these traces are connected to user profiles and user interests, they provide useful mining data that can apply to the future users which share the same preferences.

On the other side, these users might have different expectations from the result of the same query. A linguistic researcher and a plantologist researcher expect different result from the same query.

No matter how strong the semantic behind used in explaining terms, no matter how precise the data mining of the system, the best results are valued based on user profile and user interests.

User Profiling

A prerequisite for developing systems providing user interactions is to rely on user profiles. A user profile is a structured representation of the user's presence within a portal. Through this profile, the system is able to recognize each registered user. The users can interact with the system and update their personal settings. A simple profiling system is programmed in the HDL.

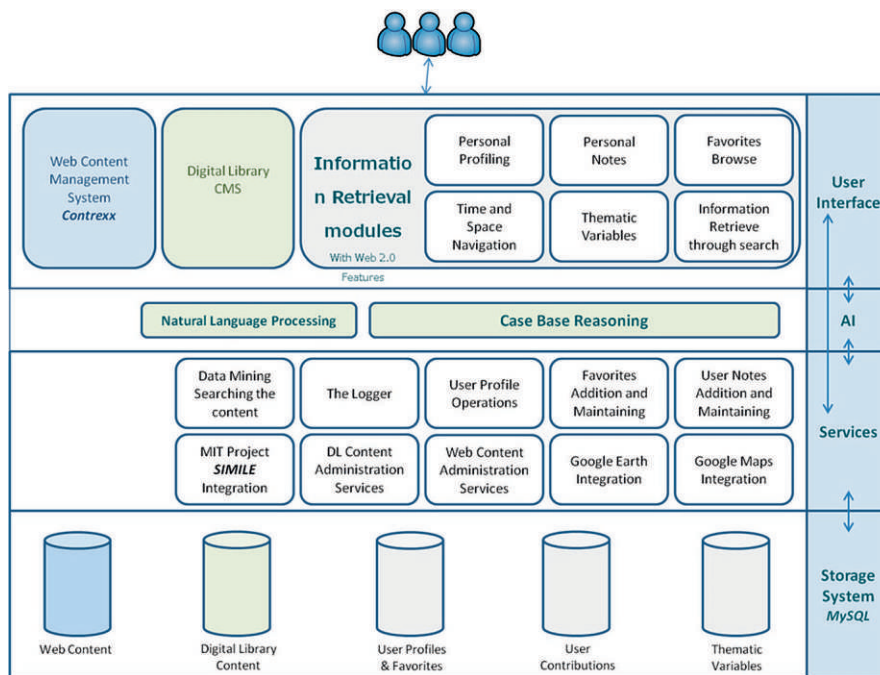


Fig. 4.4-1: General description and modules of the Humboldt's digital library and network

The user's profile in HDL includes a user's interests section. Every user is recommended and can easily update his interests. For example, an academic person with an interest in geography will get the highlighted results from a specific topic related to geography. The interests are grouped in categories. For each category, there is a list of specific interests from which the users can choose.

From a simple internet portal point-of-view, the presence of user profiles can be considered as just another feature to keep the user in the interested in the web-portal. It is a fact that systems serve better when they know their inhabitants. This is valid for the computer domain as well. If the system knows the background and interests of the users, then it can filter and provide them with some specific information. Based on the interests of the user, our system creates a set of statistical information about the paragraphs that may be of interest to the user, by analyzing on the experience of other users with a similar profile.

By means of user-profiling, personal-notes and favorite-bookmarks, the system retrieves information about the interests of each user.

In the personal profile section, the users may add information such as disciplines

of interests, general interests and regional interests. A composition of these interests provides a cosmos for each user. While users interact with the system through the content browsing, IR search tools or by writing personal notes related to any paragraph, they provide important feedback to the system. The system is basically learning what paragraphs are of interest to users of similar profile.

The interactions of the user with the system are handled (stored and analyzed) by a logger. The logger retrieves the interactions of the user with the system.

Based on these interactions, an algorithm provides for each user-profile, suggestions on the information that may better serve the users need. The logger together with an algorithm for suggestions provides the Case-Based Reasoning Engine. The CBR Engine takes in consideration: clicks (visits), user personal note, editor public notes, bookmarking of paragraph Etc.

Through the use of these variables, a better approach is provided in retrieving information related to each paragraph in the Humboldt digital library, in the same way that a well formed commentary would describe what the paragraph covers in the multidimensional space.

Through the CBR engine we have created a new approach to a suggestive system for the right content.

Since the number of the profiles in the system is still low, we are looking for new implementations of features which can make the actual users interact more with the system, and add even more valuable content to the system.

So far, the implementations in the HDL have been pointing to the creation of a research environment for the scientific users interested in the works of Alexander von Humboldt. The users of the system can make bookmarks, can write their own personal notes and in specific cases, they can write public notes related to each paragraph.

Academic Network

Many of the scientific observations made by Humboldt can be extended with additional information from academic researchers. Journals and other proceedings have a high expectation and acceptance criteria. Most of the time these scientific publications are the result of a long research work which can be described in small steps.

The blog and the concept of the personal blogs have had an increasing impact in the last years in the internet. Blog and the feedback on the blog post have created a strong community connection in many fields. We have chosen to extend our Virtual Research Environment in the

HDL with the presence of a scientific blog.

Since the beginning of this year, we have opened a blogging system for all the scientific users of our academic network. The visitors don't need any advance and expertise in the internet to start posting blog posts on the HDL website. They just need to locate the blog menu and follow the instructions below each blog. Every user can easily create a post, edit post, approve or disapprove comments. All the process is run online and no additional softwares are required. The blog automatically moves the newer posts at the top, while archiving older posts by date or topic. At the moment we have started a set of default categories for the blogs. We expect the blog categories to be extended after once we receive feedback and suggestions from our users.

Through the blog system that we have introduced in the HDL, we intent to develop stronger relationship with the visitors of the HDL. The Blogs, can be an excellent way of sharing knowledge within our network.

Not yet a scientific article, each blog post can contain interesting facts explored through our library. The aim of our blog system is to open a communication and information network, which allows users to get in contact with other researchers and share notes and research results with the Humboldt community.

Conclusions

This paper describes the integration of new WEB 2.0 features to digital libraries, using the legacy of Alexander von Humboldt as an example of digital information. By using the services, which are accessible by a Web portal, researchers can work more effectively with a wider variety of primary source materials and linked data, no matter which data type or language.

By introducing Web 2.0 technology implementations, the HDL is providing access to a information network on the legacy of Alexander von Humboldt and establishing an academic collaboration platform around the library content. Further work can extend the data content of the system including the sources, the results of the research on Humboldt, and the works of other scientists from the scientific community via public notes, blogs and similar tools. Actually a remarkable number of international scientists are part of our academic network at www.avhumboldt.net

Referenzen/References

- [1] PhD Candidate Humboldt-Universität zu Berlin
- [2] Doherr D.: Forschungsbericht — Projekt "Humboldt on the Web". Technische Berichte IAF, Hochschule Offenburg, April 2008, ([http://www.avhumboldt.net/_publications/Humboldt on the Web.pdf](http://www.avhumboldt.net/_publications/Humboldt%20on%20the%20Web.pdf))
- [3] Doherr D.: The Humboldt Digital Library: Exploring Innovative Structures. — Forschung & Technik, IAF Hochschule Offenburg, Offenburg, 2006
- [4] Brahaj A.: Concepts and Implementations of a Multivariable Search Path in the Humboldt Digital Library. — Master Thesis, University of Applied Sciences Offenburg, September 2008
- [5] Doherr D., Brahaj A.: Information Management beyond Digital Libraries- PIK — Praxis der Informationsverarbeitung und Kommunikation. Band 32, Heft 3, Seiten 161–166, ISSN (Online) 1865 — 8342, ISSN (Print) 0930 — 5157, DOI: 10.1515/piko.2009.0030, Juli-September 2009

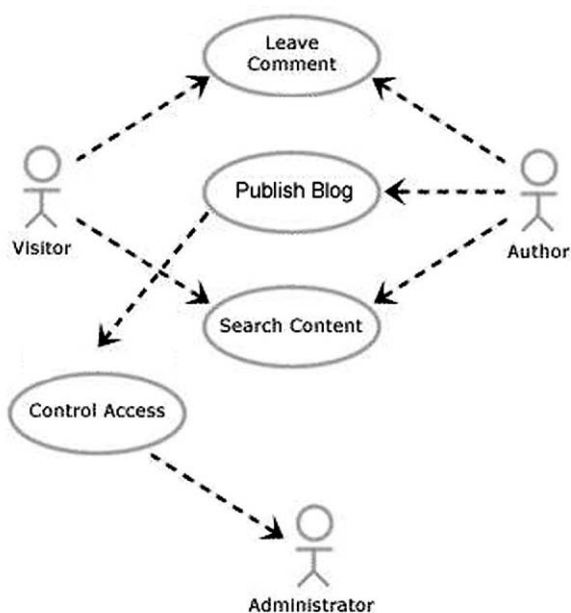


Fig. 4.4-2: User interaction with the blog system