Humboldt's Idea of Interconnectedness as an Internet Portal

Prof. Dr. Detlev Doerr
University of Applied Sciences Offenburg
Offenburg, 77652, Germany

ABSTRACT

This article sets the focus on methods of information technology in the Humboldt Portal, which represents an ongoing research project to develop a virtual research environment on the Internet for the legacy of Alexander von Humboldt.

Based on the experiences of developing and providing the Humboldt Digital Library (www.avhumboldt.net) for more than a decade, we defined a working plan to create an Internet portal for comprehensive access to Humboldt’s writings, no matter if documents are provided as PDF files, scan images or XML-TEI documents on external archives (Google Books, Internet Archive, Deutsches Textarchiv, Bibliotheque National de France).

Going far beyond services of a digital library we will provide an information network with multimedia assets, which are containing objects like terms, paragraphs, data tables, scan images, or illustrations, together with correlated properties like thematic linkage to other objects, relevant keywords with optional synonyms and dynamic hyperlinks to related translations in different languages.

So the Humboldt Portal can contribute to the key question, how to present interconnected data in an appropriate form using information technologies on the Web.

Keywords: Archives, Alexander von Humboldt, Internet portal, information network, digital library, interconnected data, dynamic hyperlinks, Google Maps, GIS

1. INTRODUCTION

The digital content of the Internet is growing exponentially and makes it possible to everyone to share information online. Even printed documents are scanned and partially converted to text by optical character recognition and republished as digital Web documents in online archives like Google Books, Internet Archive, Open Library, and others. To improve the services of digital libraries for online research, presentation and preservation, many texts are converted into a XML format, which follows the standard text representation in digital form. This standard is collectively developed by a non-profit membership organization called Text Encoding Initiative (TEI).

Alexander von Humboldt, a German scientist and explorer of the 19th century viewed the natural world holistically and described the harmony of nature among the diversity of the physical world as a conjoining between all physical disciplines. After his travels to the Americas he was admired for his ability to see the natural world and human nature in the context of a complex network of relationships that involved all known scientific and humanistic disciplines. During his South American travels, Humboldt noted in his diary: “Alles ist Wechselwirkung” everything is interconnectness and interdependent.

2. THE KEY QUESTIONS

So how can a digital library, more than a printed version of his works, represent this ambitious approach to the “unity of nature”? Is it possible to use information technology to represent the holistic character of Humboldt visions [1]? How can we compare texts in different formats, correlate translations of different paragraphs and generate a travel route with his observation points on a map? And finally how could we compare his discoveries to recent situations?
3. THE TECHNICAL APPROACH

The enormous number of digitized volumes available on the Internet makes it difficult to overview the published documents. And it isn’t possible at all to get full access to Humboldt’s idea of nature and the vision of so called “Humboldtian Science” [4] without a proper presentation of his legacy. As shown with the key questions we urgently need an information system for better understanding of the Humboldt’s works with consequences for the historic dimensions of natural sciences up to research of environmental protection and sustainability.

The first step of the development was a design of a new Internet portal, where we can provide a list of hyperlinks to relevant Internet sources as a digital library. More than only a Web server or a digital library platform, we installed a cluster of Linux servers to migrate the methods deployed in the Humboldt Digital Library [3] to our portal system.

Our starting point was the integration of TEI-conform text documents into the information network. The challenge here is the conversion of XML-files to information units and their mechanical embedding into the information network. To create the information objects we separated the information from the XML-TEI documents and imported them into a new database with a structure, comparable with the Humboldt Digital Library.

We added a Webcrawler to our portal, which is identifying all Humboldt documents on Internet and analyzing the evidence of text information and the searchability. The portal (see fig. 1) provides access to those datasets, no matter if internal or external archives are linked. All recognized documents are listed in a data table, which makes it easy to access to the relevant documents. But additionally we developed basic functions for keyword searching and analyzing evidence of text terms not only in one but all included documents.

One of the most important purposes of the portal is the visualization of Humboldt’s travels and the presentation of landmarks and locations in a map (see fig.2). As in the HDL we used the Application Programming Interface (API) of Google Maps [1]. Facing an enormous increase of data and expected requirements of data filtering we developed an XML-parser to select relevant location names and/or coordinates from the datasets and define a KML-dataset for the Google API in a dynamic way.

Together with the implemented thematic and keyword search this tool enables respond to questions, where Humboldt described plants or animals, made measurements or observed environmental processes. We added this feature to the portal and defined overlay maps, containing the images of Humboldt’s maps. In the fig.2 there is a map of Mexico, published by Humboldt, presented in Google Maps together with the digitized travel route and recognized location marks.

The portal contains an information retrieval module, which can find relevant data in external archives, the linked Humboldt Digital Library and the internal database. Additionally we designed dynamical hyperlinks to the Google Books documents, which we can provide now as document sources of Humboldt’s volumes. The information retrieval module identifies the searched document and the keyword, which is used for the searching. What we get from external archives like the Archive.org, is a complete file with highlighted keywords, but no automatic navigation to the entire

![Fig.1. Humboldt Portal as a data table on http://humboldt.hs-offenburg.de](http://humboldt.hs-offenburg.de)

![Fig.2. Humboldt’s travel route on Google Maps together with an overlay map of Mexico and identified location marks with connection to the documents in the portal.](http://humboldt.hs-offenburg.de)
paragraph in the document. This would be possible only for XML- documents or text based database access as designed in the HDL.

4. WORKING PLAN FOR AN INFORMATION NETWORK

As shown before, we migrated the methods deployed in the Humboldt Digital Library to the Humboldt Portal to provide a data table with a list of links to relevant Internet sources as digital libraries, external archives, and internally embedded XML-TEI documents.

Parallel to the text data, which is structured as a relational data table with a text object, identified by a unique primary key, there are several related metadata, containing information about images, illustrations, external links to relevant archives and automatically generated data about the use of the information retrieval module and successfully used keywords, synonyms and thematic references.

In the HDL, we implemented a method to assist the researchers finding relevant information in the digital library based on user behavior approach. The system is calculating a heat map of relevance according to the user’s profile of interest, which can be defined in the procedure of acquiring for a personal userid, as well as the behavior of other users who share similar profiles [2]. So the system can highlight paragraphs, which may be relevant to users and/or for chosen keyword search.

Unfortunately we haven’t got such a dataset in the portal because of the lack of userid and user profiles. But we are providing a complete list of relevant keywords, which will be presented in the retrieval module together with the autocomplete- function. This list of searchable terms is automatically generated from the text objects of the portal and represents the content of the Humboldt database including all recognized text documents on the Internet.

The user-behavior recommendations are assured by a case based reasoning (CBR) module in the HDL, which takes a semantic analysis of paragraphs into consideration to improve correctness and solve some problems that arise from the user-behavior approach alone [2]. We have in mind to import those features into the portal, but we are providing only text objects instead of paragraphs, which makes it complicate to implement semantic analysis.

5. TRAVELOGUES AND LOCATIONS

Many works of Alexander von Humboldt contain travelogues and location information as visited places, references and comparisons to other locations. To illustrate this graphically to the user, a dynamic Google Maps application is deployed, which represents travel routes and places from the works. The route points and places are associated with the search function of the portal. This allows the user, by clicking on a location, to perform a search for the location name in the portal (see fig. 2).

Humboldt's itineraries are known and are digitized to a vector data [4]. To extract the suspected locations of his text descriptions and their exact coordinates, all routes were first imported into a geographic information system (QGIS). Furthermore, all available worldwide place names from the geographical database GeoNames [5] were imported as well. Since the number of records is too large and the place names are not unique, a corridor of 100 km of width was defined in the GIS on both sides of the main travel route. This area determines the search area for each location (see fig 3) and significantly reduces the number of eligible place names.

Fig 3: Selected place names of the GeoNames database inside a search area of Humboldt’s travel route to the Americas

In order to improve the quality of data, the identified place names were matched against the complete term list and text objects of the entire Humboldt portal. Only those place names, which are recognized as terms in the portal, are considered as relevant information. The remaining place names are passed to the Google Maps together with the identified metadata. This is done through an XML parser, which transforms the metadata of the portal to a virtual XML file, parses and transfers into the Google Maps API. This procedure is similar to that described for the HDL [2].

Thus, the itineraries, the maps as overlays and location
descriptions together with the identified coordinates can be defined as placemarks in Google Maps. But also the inclusion of external services such as Panoramio (Photo-sharing community) as an extension of the visual presentation of Humboldts destinations in the form of roads, satellite scenery or terrain maps of Google Maps is provided [3] in the portal.

6. VISIONS

We developed the technology to create the Humboldt information network and increase the level of interconnectedness of data by user behavior and correlation of objects to other documents and even to external archives.

What is missing today, is the overall data structure of internal and external data and the worldwide valid meta data information of those information objects. This has nothing to do with TEI standards, but should describe new correlation methods and properties of information objects. We went a step forward with an information network on the Humboldt’s legacy, because this is a wonderful example to show the usability of an information network in comparison of a PDF- file service.

7. REFERENCE


