Hypertextuality in the Alexander von Humboldt Digital Library

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

and

Andreas Jankowski, M.Sc.
University of Applied Sciences Offenburg
Offenburg, 77652, Germany

ABSTRACT

To do justice to the legacy of Alexander von Humboldt, a 19th century German scientist and explorer an information and knowledge management system is required to preserve the author's original intent and promote an awareness of all his relevant works. Although all of Humboldt's works can be found on the internet as digitized papers, the complexity and internal interconnectivity of the writings is not very transparent. Humboldt's concepts of interaction cannot be adequately represented only by digitized papers or scanned documents.

The Humboldt Portal is an attempt to create a new generation of digital libraries, providing a new form of interaction and synthesis between humanistic texts and scientific observation. The digital version of his documents supplies dynamic links to sources, maps, images, graphs and relevant texts in accordance with his visions, because "*everything is interconnectedness*".

Keywords: information technology, Humboldt, portal, digital library, knowledge management, dynamic hyperlinks, interconnectedness

1. INTRODUCTION

The internet provides a comprehensive library and makes the worldwide accumulated knowledge available to users regardless of time and location. The information is accessible mostly uncensored worldwide and the scientific barriers often disappear. In the era of digital information, the trend of transforming data from analogous to digital to provide documents in digital libraries within the internet is increasing. The content of those digital archives are scanned paperwork or XML-documents, which include an internal structure of the document's text and possibly a linkage of text terms to other sources as hyperlinks.

The most popular platforms providing digital content follow the Open Archive Initiative (OAI), which "develops and promotes interoperability standards that aim to facilitate the efficient dissemination of content" [1].

The functionality of search engines generally does not go beyond textual information, since automated processes in information and knowledge recognition of valid information are missing. Books are normally digitized as static texts in digital data files, which provide searchable text information, but fail to discover logical metadata or data classifications. Today we have advanced options and attempts to provide information in digital libraries. However the technology is insufficiently developed to adequately illustrate the scientific works and the foresight of the researcher Alexander von Humboldt, a German explorer and scientist of the 19th century, who viewed nature holistically and thought an approach to science was needed to account for the harmony of nature in the diversity of the physical world. Humboldt believed in the importance of accurate measurements and precise description of observations, which should be measured by using the finest instruments and techniques available, to better detect processes and phenomena. This collected data was the basis of his scientific understanding. This quantitative methodology would become known as "Humboldtian science" [2]. However, it is difficult to find relevant information in digital archives without a clue of text terms, phrases, or a specific title, e.g. a Latin plant name. And it is nearly impossible to find comprehensive context-related data or search successfully for different topics within digital libraries, such as searching for a complete list of locations Humboldt visited or Humboldt's detailed data about air pollution.

This was the logic behind developing the Humboldt Digital Library (HDL) [3] which was created at the University of Applied Sciences Offenburg in cooperation with the University of Kansas, Lawrence, and has been online for many years now. It grants free access to digitized documents from Alexander von Humboldt [4]. New developments of the Humboldt Portal open access to all of Humboldt's volumes in different languages in digitized forms or as links to external archives [5].

The vision of Humboldt demands a quantum leap of knowledge management and semantic web structures, which could lead to the kind of information system Steward Brand had in mind: "*The future libraries should protect our 'right to know' and 'right to remember', what means, that the Internet archive now is 'the beginning of a cure- the beginning of complete, detailed, accessible, searchable memory for society*" [Stewart Brand, president of The Long Now Foundation]

2. Humboldt's Vision of Interconnectedness and Hypertextuality

Humboldt's aspirations to observe all aspects of nature within the context of its interactions and dynamic processes, makes it insufficient for our understanding to simply try and learn from digitized versions of his work. . With his methods, Humboldt laid the very foundation for the modern vision of sustainability and sustainable development. [5]

Thus forming the very modern concept, Lucht reported 2009: "*Alexander von Humboldt asked himself among other things, how an earth science that shows the interaction of the earth, the life and the human being has to look like"* [6].

Humboldt held the strong conviction that the distinct academic disciplines represented artificial divisions of knowledge [2] and the key to understanding nature was interconnectedness. During his South American travels, Humboldt noted in his diary: "*Alles ist Wechselwirkung":* everything is interconnected and interdependent.

As a result, he published a detailed narrative of his travels, constantly integrating his observations and data from various perspectives and disciplines. The need to make connections was characteristic of his publications.

The complexity and interconnectedness of information in his works are neither visible in his digitized writings nor in document-oriented digitized content. Humboldt's emphasis on interconnectedness can sometimes be gauged by his images and drawings, which document natural processes and correlations, which therefore means Humboldt's concept of interactions can only be modelled and traced adequately with the help of modern information technologies [4].

3. Working Plan for A Dynamic Knowledge Base

We defined a working plan to create an internet portal for comprehensive access to Humboldt's writings, regardless of the documents' digitized format (PDF files, scanned images, or XML-TEI documents on external archives such as Google Books, Internet Archive, Deutsches Textarchiv, Bibliotheque National de France). The portal contains an information retrieval module, which can find relevant data in the Humboldt Digital Library and the internal XML- database. Additionally, we designed dynamic hyperlinks to Google Books documents, which we can now provide as document sources of Humboldt's volumes. The information retrieval module identifies the searched document and the keyword which is used. We then receive a complete file from Google Books with highlighted keywords but no automatic navigation to the entire paragraph in the document. This would be possible only for XML- documents or text based database access as designed in the HDL.

This service goes far beyond online services of digital libraries. But we have concluded that the aim of the development must be the provision of dynamic information systems and knowledge networks, in which information is collected by the computer system and made available depending on the search request [7].

To meet the referred requirements, it is necessary to handle a search request in the entire context, which is defined by metadata such as origin, thematic embedding, environment, geographical position, etc. Instead of search terms, we are dealing with objects which can be formally named and classified by computer processes defining ontological types, properties, and interrelationships. In addition, logical connections like synonymous terms and hyperlinks to comparable internet information must be correlated to those objects, too. We define dynamic hyperlinks as flexible, at runtime generated objects. Whereas the standard definition for a hyperlink is an object that consist of properties like the address of the destination and the parameters that affect its behavior and is limited by the external archive functionality. The building process of a hyperlink object depends on which function will be used, e.g. searching or only viewing a document. In the case of a search in an external archive, an additional keyword has to be provided. This is the simplest version of a dynamic hyperlink.

With this starting point, an extended dynamic hyperlinked object is supplemented by environmental information. That information can be the source from where the search is executed or in what context an object is generated. As a result, the destination can react to that and present a dynamic output. The result of a map-generated search should differ from the search in a document using text terms or keywords. In the first case the result has to be in relation to geographical matter. The other search may lead to different outputs based on the starting point.

These kinds of dynamic hyperlinks can only work in an environment which provides the required interfaces for dynamic hyperlinks.

4. TECHNICAL APPROACH

What is the difference between the Humboldt Digital Library and a digital library as an online archive? The HDL doesn't only provide digitized documents such as PDF-files or images from book scans, but also text information and metadata as object information embedded in an information network. Therefore the system is representing an information system rather than a digital library, which is also reflected by the implemented functions for searching for texts in different versions and translations, comparisons of paragraphs of different documents, and presentation of images in their contexts.

The Humboldt Portal [9] takes a different approach in terms of data storage, and offers a collection of external digital works of Alexander von Humboldt. The aim of the portal is not to reinvent the wheel of a digital library with its internal data storage, but more the building of a linked information network which consists of different information sources and archives. Initial success has already been gained with the implementation of an archive cross-search feature. The usability of the search engine is supported by an automatically generated tag list for an auto-complete function that provides suggestions for related terms while entering the keyword. The search results are listed together with supplemented links to the original sources, which enables rapid access to the search functions of the linked external archive. Depending on the available interfaces of external archives, the search term can be directly transferred to start a search in the original document.

As explained in a former article [5]., we embedded the concept of dynamic hyperlinks into the library, based on the individual paragraphs in the Humboldt works in the form of media assets, which enables the use of the Application Programming Interface (API) of Google Maps for geographical as well as text content navigation.

To improve the geographical feature, the names of many places in the works of Alexander von Humboldt were semi-automatically extracted and linked to Google Maps. In this process, the content of his writings was checked against a location database (GeoNames [8]) and relevant information tables were automatically generated. Those tables are used to create a context-sensitive search for locations correlated to the travel routes of Humboldt or those approximate to that.

Those concepts are implemented in the analysis and data processing of location names in Humboldt's works (see fig. 1), which were automatically detected from the HDL database in connection with other internal and external information. The expansion of the text objects by integration with location data represents a feature of the interactive map implementation with the Google Maps API.

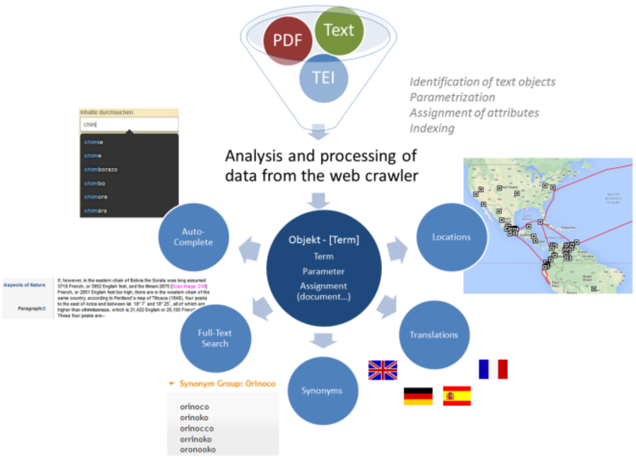
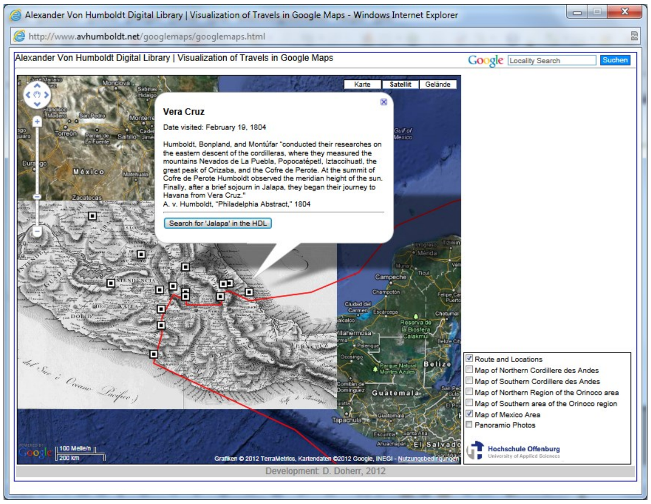


Fig. 1: Analysis and data processing of data that were integrated into the portal via web crawler. The terms were converted to text objects, which then were expanded by parameters. The parameters are used for the auto-complete function to find synonyms, to identify translations, and to assign locations.

Among other features, the extracted location objects can be viewed in the geographical context, and every location is presented with a dynamic link, so a search can be started directly out of the map. An additional presentation of original maps allows a more detailed view.


Alexander Von Humboldt Digital Library | Visualization of Travels in Google Maps - Windows Internet Explorer

http://www.avhumboldt.net/googlemaps/googlemaps.html

Alexander Von Humboldt Digital Library | Visualization of Travels in Google Maps

Vera Cruz

Date visited: February 19, 1804

Humboldt, Bonpland, and Montúfar "conducted their researches on the eastern descent of the cordilleras, where they measured the mountains Nevados de La Puebla, Popocatépetl, Iztaccihuatl, the great peak of Orizaba, and the Cofre de Perote. At the summit of Cofre de Perote Humboldt observed the meridian height of the sun. Finally, after a brief sojourn in Jalapa, they began their journey to Havana from Vera Cruz."
A. v. Humboldt, "Philadelphia Abstract," 1804

Search for 'Jalapa' in the HDL

☑ Route and Locations
☐ Map of Northern Cordillere des Andes
☐ Map of Southern Cordillere des Andes
☐ Map of Northern Region of the Orinoco area
☐ Map of Southern area of the Orinoco region
☑ Map of Mexico Area
☐ Panoramio Photos

Grafiken © 2012 TerraMetrics, Kartendaten ©2012 Google, INEGI - Nutzungsbedingungen

Development: D. Doherr, 2012

Fig. 2: Google Maps application of the HDL by using Humboldt's accurate maps as Google overlay maps and place marks, which are digitized or generated by Humboldt's descriptions and geographical coordinates, coming from the GeoNames database.

By using the Google Maps features of the HDL, the user will find several locations which are marked by place marks and contain descriptions and chronology of when Humboldt visited this location. As a connection to the HDL, every place mark contains a link button which opens a browser window with an online search for paragraphs where Humboldt described that location or where the user could find more details about his observations (see fig.2).

We defined maps from Humboldt as overlays to the recent satellite scenery of Google and georeferenced them as well as possible. The overlay maps can be used to find locations (regardless of the identical place names), to identify geographic observations of Humboldt, and compare landscapes from 200 years ago to more recent ones.

To improve the internet services and get closer to the true visions of Humboldt, we used the object oriented structure of the text documents to create more dynamic hyperlinks to other sources, which could provide potentially relevant information.

One of the external sources is a knowledge based information system "Wolfram Alpha" [10]. It is an online service developed by Wolfram Research that answers factual queries directly by computing structured data. In comparison to Google, the results of the request are presented in a predefined form, which depends on the recognized form of data structure such as object properties.

It is built on top of Wolfram's Mathematica, computing software for computer algebra, symbolic and numerical computation, visualization, and statistics [10].

Much like the aims of our development of the knowledge based Humboldt Digital Library, the developers of Wolfram express, that "*one goal of the project is to collect and curate all objective data; implement every known model, method, and algorithm; and make it possible to compute whatever can be computed about anything. Another goal is to build on the achievements of science and other systematizations of knowledge to provide a single source that can be relied on by everyone for definitive answers to factual queries.*" [10]

Because of the availability of the Wolfram Alpha API, it is possible to implement the service into our own developments and take advantage of its capabilities, as shown in fig.3. This opens the option to automatic classification of text of all documents, which we included in our portal. Whenever a user of the Humboldt Portal inserts a term into the search engine, first the system is checking for occurrences inside the text of the internal and external documents. Secondly a dynamic hyperlink will be created to access to the external archive, which was identified as a source of the found document. Our concept includes a generalization/ specialization of the search term related to the GeoNames database and the Wolfram Alpha, to classify the term, if possible.

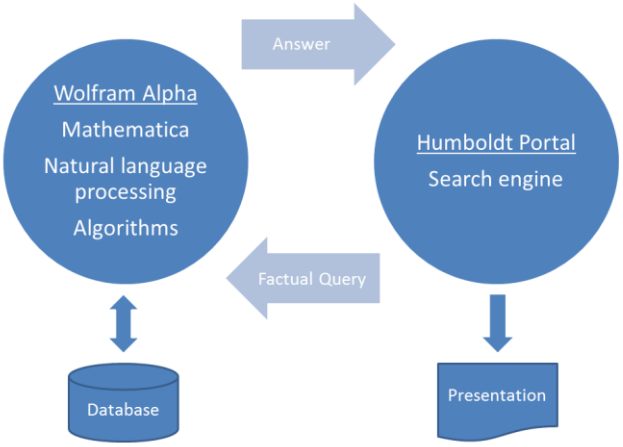


Fig. 3: Information cycle of the Wolfram Alpha implementation. The search engine of the Humboldt Portal generates and sends a query to Wolfram Alphas's knowledge engine. Wolfram Alpha computes the query with information stored in a database and returns the answer to the portal. Finally, the portal search engine presents the result to the user.

The importance of that development is that we have all relevant text terms, notations, measurements, and observations which Humboldt ever described in his works at our fingertips. The answer from the Humboldt Portal to a search request now can be presented as a list of documents from internal or dynamical linked external archives, which are containing the term as text information. If the automated classification process of the Humboldt Portal recognizes a classified object, the output can be different and possibly will lead to additional information. This concept is what we call "Hypertextuality".

Whether searching in a location- and GeoName-database, or for occurrences of the same term in different volumes, or for searches in the Wolfram knowledge base, the request will lead to a result, which can be presented by text, Google Maps place marks, or an object description, depending on the results of Wolfram's search. Based on this concept of "Hypertextuality" the Humboldt Portal can be used as an example for interconnected hyper-searching.

5. SUMMARY

We implemented the methods used in the Humboldt Digital Library in a new internet portal, which provides a list of links to relevant internet sources as a digital library, embedded XML-TEI documents and the HDL-documents. Additionally, the dynamic hyperlinks from the HDL to other online archives and the connection to Wolframs knowledge engine open innovative options for better understanding Humboldt's visions of nature.

The coming challenges may be a more dynamic hyperlink generation, the decoding of multilingual text terms, the automatic generation of ontological structures, and the decoding of images and data tables to better get the computer to understand what we are thinking. This is exactly the path to artificial intelligence.

6. REFERENCE

[1] **Open Archives Initiative:** http://www.openarchives.org (access: 27.08.2015)

[2]Wikipedia, **Humboldtian science** https://en.wikipedia.org/wiki/Humboldtian_science (access: 27.08.2015)

[3] **Humboldt Digital Library:** http://www.avhumboldt.net (access: 08. Febr. 2016)

[4] D. Doherr & F. Baron, **Humboldt Digital Library and Interconnectedness,** The Environmentalist, ISSN 0251-1088, DOI 10.1007/s10669-011-9369-y, Springer-Verlag, 2011

[5] D. Doherr, **Interconnectedness und digitale Texte**, HiN - Humboldt im Netz, Internationale Zeitschrift für Humboldt-Studien XIV, 26, S. 12-18. , Potsdam 2013. www.uni-potsdam.de/u/romanistik/humboldt/hin/hin26/doherr.htm

[6] W. Lucht, **Lecture at HU Berlin on sustainability science**, Download under: http://www.hu-

berlin.de/pr/medien/publikationen/humboldt/2008/200905/humboldt_200905.pdf, 2009

[7] A. Brahaj, D. Doherr, J. Hoxha, **Behavior-Based Information Seeking in Digital Libraries,** 2. International Multi-Conference on Complexity, Informatics and Cybernetics IMCIC, Orlando, 2011, USA

[8] **GeoNames**
http://www.geonames.org (access: 08. Febr. 2016)

[9] **Portal Alexander von Humboldt**,
http://humboldt.hs-offenburg.de (access: 08. Febr. 2016)

[10] NIDHI S, **Wolfram|Alpha: A Computational Knowledge Engine**, Seminar Report, Cochin Univ. of Science & Technology, 2011, India, dspace.cusat.ac.in/xmlui/handle/123456789/3503