The Humboldt Portal: Complexity and Interconnectedness

Detlev Doherr

University of Applied Sciences
Offenburg, Germany

Abstract

The Humboldt Portal has been designed and implemented as part of an ongoing research project to develop an information system on the Internet to share the documents and rare books of Alexander von Humboldt, a 19th century German scientist and explorer, who viewed the natural world holistically and described the harmony of nature among the diversity of the physical world. Even after more than two centuries he is admired for his ability to see the natural world and human nature in the context of a complex network of relationships. The design and implementation of the Humboldt Portal are also oriented to support further research on Humboldt's intellectual perspective.

Although all of Humboldt's works can be found on the internet as digitized documents, the complexity and internal inter-connectivity of his vision of nature cannot be adequately represented only by digitized papers or scanned documents in digital libraries.

As a consequence a specific portal of the Humboldt's documents was developed, which extends the standards of digital libraries and offers a technical approach for the adequate presentation of highly interconnected data.

Due to the continuous scientific and literary research, new insights and requirements for the digital presentation of Humboldt documents are constantly emerging, so that this article only provides a summary of the concepts realized at now. Consequently, the design and implementation of the Humboldt Portal is both: a consequence of a continuing research project and oriented to support more research on Humboldt´s intellectual holistic perspective, which was an anticipation to the System Approach of the last Century.

Keywords: Digital Library, Alexander von Humboldt, Internet Portal, Interconnected Data

1. INTRODUCTION

The digital content of the Internet is growing exponentially, enabling everyone to share documents and rare books. Printed documents are scanned and converted into text information by optical character recognition and republished as digital web documents in online archives and digital libraries[1]. To improve the services of digital libraries for

online research, presentation and preservation, many Humboldt texts are converted into an XML format that meets the standards of text representation in digital form developed by an organization called the Text Encoding Initiative (TEI) [2].

Even after more than two hundred years, Alexander von Humboldt is still admired for his ability to see the natural world and human nature in the context of a complex network of relationships [3]. After his travels to America, he was admired for his vision of nature, where the natural world and human nature are linked in the context of a complex network of scientific and humanistic disciplines. During his journeys to South America, Humboldt noted in his diary: "Everything is interaction" everything is interconnected and interdependent [1].

The complexity and interconnectedness of the information in his works, however, are not visible either in his writings or in document-oriented digitized content [4]. His knowledge of the complex interconnectedness of nature as a result of natural processes can only be detected in his illustrations and drawings.

2. THE KEY QUESTION

In the 19th century Humboldt explored nature and "conceived a bold new vision of nature that still today influences the way we understand the new world" [5]. He crossed the boundaries of the known world, traveled to America [6] and the outermost parts of Siberia, and developed concepts for a scientific approach to science. He used his procedural thinking to relate his observations to each other and to place the unknown within his vision of nature. He described his view of nature, where he found connections everywhere, even between animated and inanimate nature. In this large chain of interrelated effects, he noticed that no single fact can be considered in isolation. "So he invented the web of life as we know it today"[5].

For two centuries scientists have been working on the interpretation of Humboldt's heritage, described in detail in the documents, but have difficulty in presenting a complete network of interdependent and multivariate data that could adequately represent Humboldt's vision of nature as a fascinating harmony of complexity[1].

His books do not adequately reflect this complexity: they resemble more the print out of a web page that contains many texts and images, but lost any form of hyperlinks. Also digital libraries usually show only the scanned or textually prepared information, also the search engines limit themselves to text information, because automated processes of information processing and knowledge recognition are missing. Books offer searchable text information, but there is no way to find logical metadata or data classifications[7].

So the question comes up, what kind of information technology is necessary to represent the holistic character of Humboldt's idea of nature[3], considering the different formats of his documents as text books, illustrations, data tables of accurate measurements, drawings and handwritten diaries and travelogues. 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 modeled and traced adequately with the help of modern information technologies [4]. Consequently, the design and implementation of the Humboldt Portal a consequence of a continuing research project, started in the year 2001 [8] and oriented to support more research on Humboldt´s intellectual holistic perspective, which was an anticipation to the system approach of the last century.

3. THE TECHNICAL APPROACH

Various works of Alexander von Humboldt are available in various online archives and digital libraries. The documents are available in different qualities, translations and partly incomplete. The 'Humboldt Digital Library' [9] was developed in order to provide an overview of the collected works of Humboldt and a complete bibliography. It also provided a development environment for the IT processing of the networked data. Therefore, all Humboldt documents were identified on the Internet and analyzed on evidence of text information and the searchability. We set focus only to reliable sources, because the quality of the documents is much more important than the number of re-published documents [1].

On the basis of the 'Humboldt Digital Library' the 'Humboldt Portal' [10] was developed, which is offers a different approach in terms of data storage and collection of external digital works of 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 [7].

The portal (see figure 1) provides access to these data records, regardless of whether internal or external archives are linked. All recognized documents are listed in a data table, which allows easy access to the relevant documents. In addition, we have also developed basic functions for keyword search and analysis of evidence for text terms not only in one but in all integrated documents.

In order to integrate the documents to the portal, the available TEI- compliant text documents were integrated into the information network, whereby the XML files had to be converted into information units and imported into a new database with a structure comparable to the 'Humboldt Digital Library' [1].

The portal contains search functions both for the internal text objects and for all external archive documents. Dynamic hyperlinks are generated for these search functions. These hyperlinks consist of the text information and the Internet address of the external document. The building process of a hyperlink object also depends on which function is required in the external archive, for example, search or display of a document. These parameters influence the reactions of the external archive, but are restricted the external archive functionality [1].

Figure 1: Humboldt Portal with Data Table and Listed Documents [10]

One of the most important purposes of the portal is the visualization of Humboldt's travels and the representation of locations and places on a geographical map (see figure 2). As in the 'Humboldt Digital Library' we used the Application Programming Interface (API) of Google Maps [3]. In view of the enormous data growth and the expected data

filtering requirements, we developed an XML parser to select relevant place names and/or coordinates from the datasets and define a KML dataset for the Google API [1].

To improve the geographical feature, the names of many places in Humboldt's documents were extracted semi-automatically and linked to Google Maps [7]. The content of his writings was compared with a location database [9] and relevant information tables were automatically generated. From these tables, a context-sensitive search for places correlating with Humboldt's well-known itinerary is created. These concepts are also implemented in the portal to show location names in Humboldt's works (see figure 2).

After the development of the 'Humboldt Digital Library' containing the visualization of geographical information via Google Earth, we decided to migrate this feature to the portal as well on the basis of Google Maps. Several maps were generated from Humboldt's drawings and were integrated as overlay maps. Figure 2 shows a map of Mexico in Google Maps with the digitized itinerary and location markers. The overlay maps can be used to find locations (regardless of the identical place names), to identify Humboldt's geographic observations, and compare landscapes from 200 years ago to more recent ones [1].

Some work was required to adjust the hand-drawn maps from the Humboldt documents to the exact geographic position. We obtained the necessary coordinates by comparing known locations of the Humboldt maps with today's 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.

By using Google Maps, the user has access to visualized locations identified by place marks. These locations refer to descriptions of Humboldt who visited these locations and recorded observations there [7]. Each location marker contains a hyperlink that triggers an online search in the 'Humboldt Digital Library' to refer to the descriptive text paragraph in one of Humboldt's documents. This makes it possible to search for relevant text paragraphs in all collected text documents directly from the map, which can also be supplemented by an additional display of the original maps.

A cross-document keyword search complements the portal function and is useful for thematic search. Initial successes have already been achieved with the Information Retrieval Module of the 'Humboldt Digital Library', which identifies the document sought using the keyword used for the search [1].
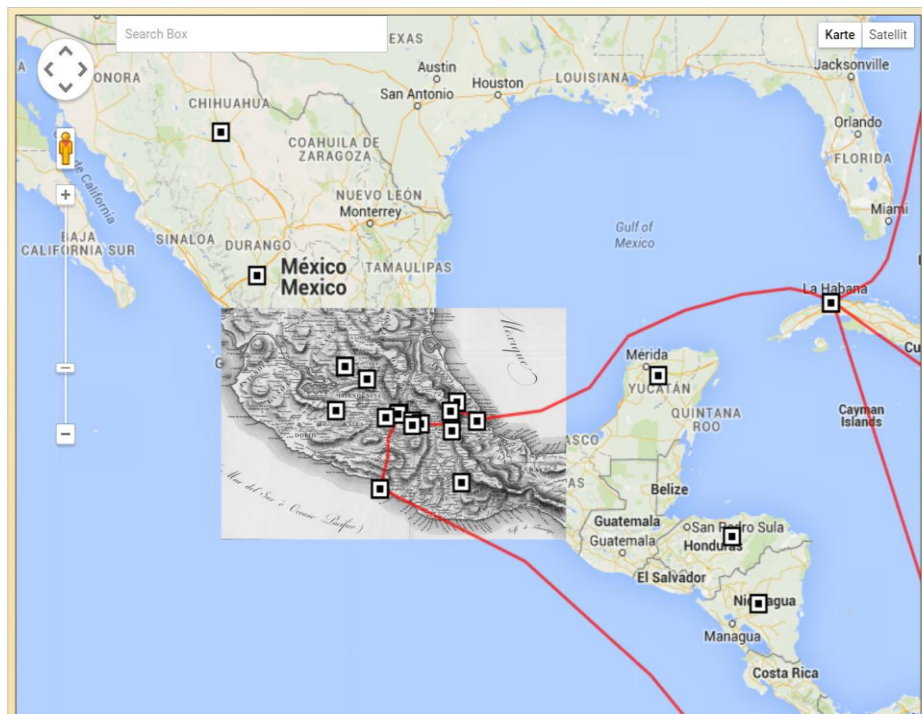
**Figure 2: Humboldt's Travel Route on Google Maps together with a
Map of Mexico and Identified Location Marks with Link to the Documents in the
Portal** [10]

For the implementation of a cross-document search in the portal, the Information Retrieval Module was extended to include the relevant data in external archives, the connected 'Humboldt Digital Library' and the internal database. The basis for this is a web crawler that analyses the linked documents, filters out the required text information and stores it in a database together with the parameters required for the dynamic hyperlink. These parameters include the Internet address, the document name, the identifier in the respective external archive and all other criteria required to find the text passage.

The search engine is integrating an automatically generated keyword list with auto-completion that provides recommendations for related terms when inserting the keyword. The search results are listed along with additional links to the original sources and providing direct access to the search functions of the linked external archive. Depending on the available interfaces of external archives, the search can be performed directly. What we get from some external archives, is a complete file with highlighted keywords, but no automatic navigation to the entire paragraph in the document.

4. CONCEPT FOR AN INFORMATION NETWORK

As already shown, we have migrated the methods used in the former 'Humboldt Digital Library' to the 'Humboldt Portal' to provide a data table with a list of links to relevant Internet sources such as digital libraries, external archives and internally embedded XML-TEI documents[1].

Parallel to the text data, which are structured as a text object, there are several associated metadata, which contain information about images, illustrations, external links to relevant archives. These metadata are automatically generated data, coming from the Information Retrieval module and successfully used keywords, synonyms, and thematic references.

We have implemented a method to support researchers find relevant information in the portal based on the user behavior approach in the 'Humboldt Digital Library'. We have not planned this for the portal, since the scientific work with the Humboldt texts must be based not only on the textual representations, but also to a large extent on the drawings and illustrations, including the handwritten notes. For these multivariate and multi-thematic requirements, there are still no solutions available in the portal, so that an automated user support would have to run into the void because of the incompleteness of the available information.

In the portal there is a complete list of the terms used by Humboldt, which can be used as relevant keywords in the retrieval module and by auto-completion. 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 text documents recognized on the Internet[1]. This is a first approach for a comprehensive information network which connects all available information with each other and also brings semantic developments into view.

5. TRAVELOGUES AND LOCATIONS

Many of Humboldt's works contain travelogues and location information such as places visited, references and comparisons with other places. To graphically illustrate this to the user, a dynamic Google Maps application is provided that displays travel routes and locations from the work. Route points and locations are associated with the portal's search function[1]. This allows the user to search for the location name in the portal by clicking on a location (see figure 2). But not all locations are already known, where Humboldt carried out his observations, measurements or collected samples . The place names of that time do not necessarily correspond to today's names or were not mentioned by Humboldt.

For this reason, a tool was developed which marks all possible locations on the basis of Google Maps depending on the distance of the travel route on the basis of Humboldt's travel routes and the digitized vector data [9]. In order to determine the assumed positions of his text descriptions and their exact coordinates, all routes were first imported into a Geographic Information System (GIS). In addition, all place names available worldwide were imported from the GeoNames [12] database. Since the number of data sets is too large and the place names are not unique, a corridor of 100 km width was defined on both sides of the main travel route in the GIS. This range determines the search range for each location (see figure 3) and significantly reduces the number of allowed place names.

In order to improve data quality, the identified place names were compared with the complete terminology list and the text objects of the entire Humboldt Portal. Only the place names that are recognized as terms in the portal are regarded as relevant information. The remaining place names are transferred to Google Maps together with the transfers to the Google Maps API. This process is similar to the one described for the identified metadata in the 'Humboldt Digital Library' [11]. This is done using an XML parser that converts the portal's metadata into a virtual XML file, analyzes it, and then sends it to Google Maps.

Thus, the travel routes, the maps as overlays and location descriptions together with the identified coordinates can be defined as place marks in Google Maps. But also the integration of external services such as Panoramio (Photo-Sharing-Community) as an extension of the visual representation of Humboldt destinations in the form of streets, satellite images or terrain maps of Google Maps is provided in the portal [9].

6. CONCLUSIONS

We have developed technical foundations to create a information network for Humboldt and to increase the degree of networking of data and correlation of objects with other documents and even with external archives.

What is missing today is the overall data structure of the internal and external data and the globally valid metadata of these information objects. This has nothing to do with TEI standards, but should describe new correlation methods and characteristics of information objects. With an information network on the heritage of Humboldt, we can take a step forward in the history of natural sciences and the Humboldt research and present a new type of digital library in approach in which the advantages of an information network become obvious compared to PDF file services [1].

Thus, the developments of the portal can be used to derive various requirements for the adequate presentation of documents in archives and, in particular, semantically networked information when it comes to generating dynamic hyperlinks, recognizing semantic structures in one or more correlated documents or forming ontological structures.
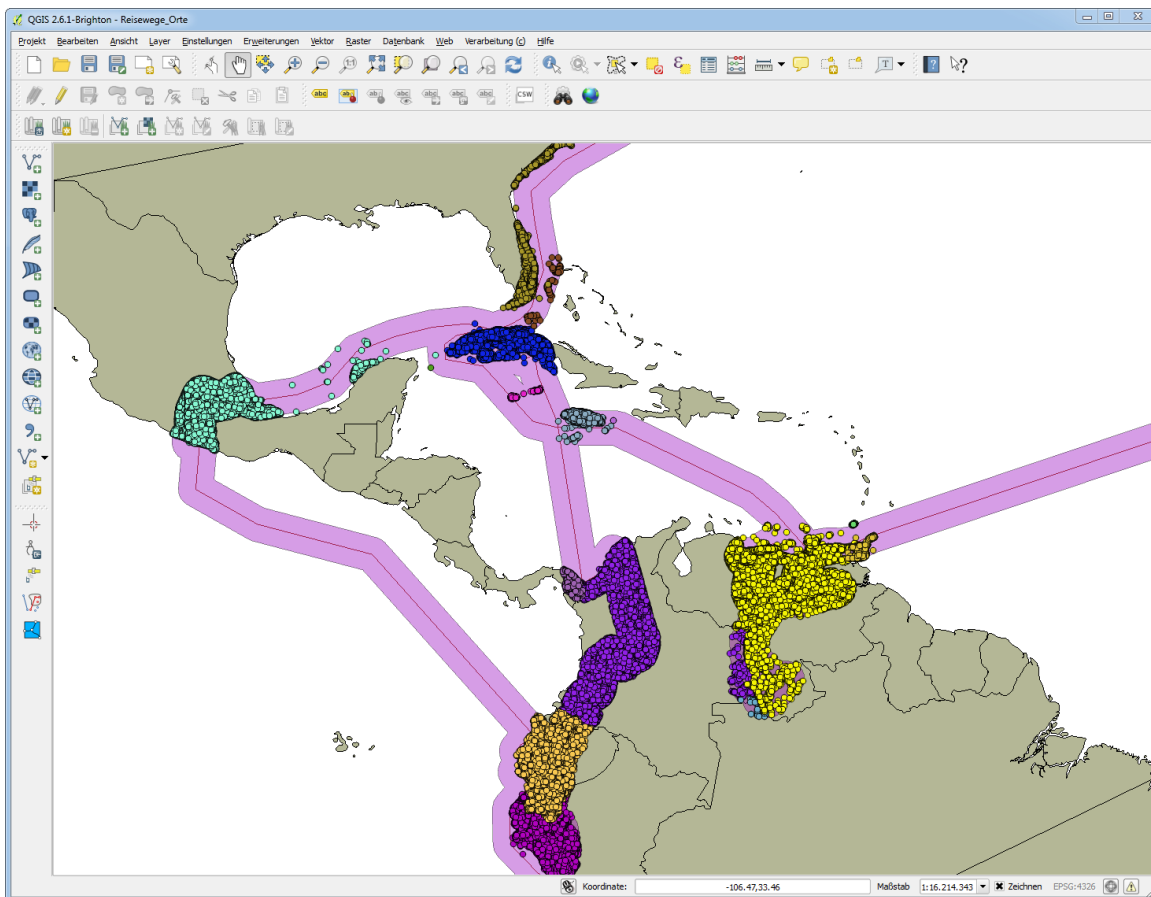

Figure3: Selected Place Names of the GeoName Database Inside a Search Area of Humboldt's Travel Route to the Americas [1]

7. REFERENCES

[1] **Doherr, D. (2015):** 'Humboldt's Idea of Interconnectedness as an Internet Portal'.- Conference: 6th International Multi-Conference on COMPLEXITY, INFORMATICS AND CYBERNETICS, Volume: Proceedings, Vol I, P.43-47, Orlando, Florida, USA

[2] **Text Encoding Initiative**. (2019, January 14). Retreived from HTTP://WWW.TEI-C.ORG

[3] Doherr, D. (2013): 'Interconnectedness und digitale Texte'.**-** HiN - Humboldt im Netz,Potsdam – Berlin XIV, 26, S. 12-18. , Potsdam 2013

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

[5] Wulf, A. (2016): 'The invention of Nature'.- 473 p., ISBN: 9781848549005, Clays Ltd. St Ives, 2016

[6] DE HUMBOLDT, A. & BONPLAND, A. (1818): 'personal narrative of travels to the equinoctial regions of the new continent '. london, 1818. retreived from https://archive.org

[7] Doherr, D. & Jankowski, A. (2016): 'Hypertextuality in the Alexander von Humboldt'.- The Journal on Systemics, Cybernetics and Informatics (JSCI), Volume 14 - Number 2 - Year 2016, ISSN: 1690-4524

[8] Doherr, D. (2005): ' The Humboldt Digital Library: Exploring Innovative Structures'.- HIN, Alexander von Humboldt im Netz, ISSN 1617-5239, Sondernummer VI, 10, Potsdam, 2005, https://www.uni-potsdam.de/romanistik/hin/hin_10.htm

[9] The Humboldt Digital Library. (2019, January 12**).** Retrieved from http://www.avhumboldt.net

[10] Portal Alexander von Humboldt. (2019, January 12). Retrieved from http://humboldt.hs-offenburg.de

[11] Brahaj, A., Doherr, D. & Hoxha, J. (2011): 'Behavior-Based Information Seeking in Digital Libraries'.- 2. International Multi-Conference on Complexity, Informatics and Cybernetics IMCIC, Orlando, 2011, USA

[12] **GeoNames.** (2019, January 12). Retrieved from http://www.geonames.org