Automatic Identification of Travel Locations in Rare Books - Object Oriented Information Management

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

The digital content of the Internet is growing exponentially and mass digitization of printed media opens access to literature, in particular the genre of travel literature from the 18th and 19th century, which consists of diaries or travel books describing routes, observations or inspirations. The identification of described locations in the digital text is a long-standing challenge which requires information technology to supply dynamic links to sources by new forms of interaction and synthesis between humanistic texts and scientific observations.

Using object oriented information technology, a prototype of a software tool is developed which makes it possible to automatically identify geographic locations and travel routes mentioned in rare books. The information objects contain properties such as names and classification codes for populated places, streams, mountains and regions. Together, with the latitudes and longitudes of every single location, it is possible to geo-reference this information in order that all processed and filtered datasets can be displayed by a map application. This method has already been used in the Humboldt Digital Library to present Alexander von Humboldt's maps and was tested in a case study to prove the correctness and reliability of the automatic identification of locations based on the work of Alexander von Humboldt and Johann Wolfgang von Goethe.

The results reveal numerous errors due to misspellings, change of location names, equality of terms and location names. But on the other hand it becomes very clear that results of the automatic object detection and recognition can be improved by error-free and comprehensive sources. As a result an increase in quality and usability of the service can be expected, accompanied by more options to detect unknown locations in the descriptions of rare books.

Keywords:
information technology, Humboldt, digital library, travel literature, Google Maps, knowledge management, interconnectedness, Alexander von Humboldt, Johann Wolfgang von Goethe.

1. INTRODUCTION

The digital content of the Internet is growing exponentially and makes it possible for everyone to share online information. Mass digitization of printed media has begun, which is the conversion of whole libraries without making a selection of individual materials [1]. The key questions are, how those digitized books can be used and serve library users?

Especially the genre of travel literature from the 18th and 19th century consists of diaries or travel books describing routes, observations or inspirations. Data tables, maps, and images contain geographic locations, travel routes and observations, which cannot be properly detected by the digitization processes. The identification of described locations in the digital text is a long-standing challenge and is made difficult due to misspellings, absent or inaccurate geographic coordinates, changes of local names, and unevaluated information on maps, which requires information technology to supply dynamic links to sources by new forms of interaction and synthesis between humanistic texts and scientific observations [2].

2. RECOGNITION OF GEOGRAPHIC LOCATIONS

The mass digitization requires a specific tool for the decoding and recognition process of geographic information in rare books because it improves the usability of the digitized documents on basis of an information system. This process should be coordinated

and controlled automatically by the digitization process to serve the users of a library and to provide a simplified geographic map containing mentioned locations of rare books.

Using object oriented information technology, an automated system is developed. This is a prototype of a software tool for information management and interconnection of datasets which makes it possible to automatically identify geographic locations and travel routes mentioned in rare books. As a first step on the automated process the program parses and transforms the text information into information datasets as implemented in the concept of the Humboldt Digital Library [4]. The next step reduces the harvested raw data to only georeferenced and classified location names using accessible online dictionaries. Those datasets contain properties such as names and classification codes for populated places, streams, mountains and regions. Together with the latitudes and longitudes of every single location, it is possible to geo-reference this information and display the locations on a map.

3. WORKFLOW - OVERALL PROCESS CHAIN

The first defined step of the process chain mentioned in Figure 1 is to analyze and to prepare the rare books from Goethe (See [3]) and Humboldt (See [4]) in different formats like TXT-, XML- or PDF-files to ensure a standardized output for further processing. The built in mechanisms ensure a reduction and elimination of irrelevant text information and generate a list of unique keywords. Afterwards, the keyword list, as the basis for further processes, is transferred to a temporary database.

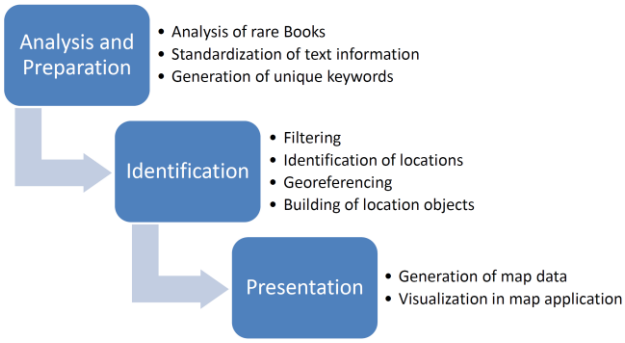

Figure 1: Process to analyze, identify , and present geographic locations mentioned in rare books

The preparation of the text information includes standardization of terms and generation of unique keyword lists. Additionally, an object oriented structure is created by using those terms, requiring more properties such as which kind of information, region, geographic coordinates, and relation to other objects. To get further details about the object, the identification process shown in figure 2 uses data mining features, which are currently limited to only some reliable data services on the Internet.

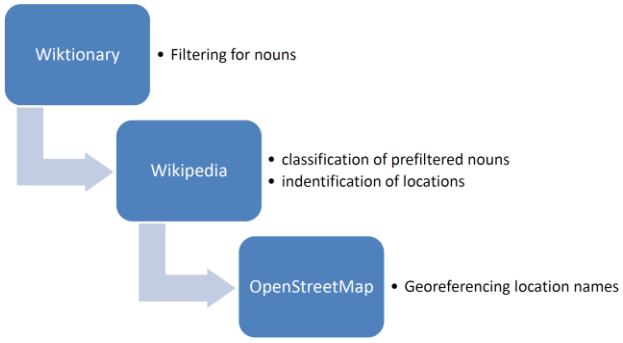

Figure 2: The identification process includes data mining features, limited to reliable sources.

The step of identification is using open source databases to filter and identify location data. Involved databases are Wiktionary, Wikipedia and OpenStreetMap. Wiktionary is used to filter out verbs and adjectives to reduce data traffic, processing time and quality of the further classification of keywords by Wikipedia. At this point, the first information objects are generated and transferred to the main database. By using Wikipedia's API, it is possible to identify whether the keywords are locations or have other classifications, enrich the related information object with properties and make further decisions depending on the result. In the case that the keyword is classified as a location, the keyword is georeferenced by an OpenStreetMap API implementation and the properties for geographic coordinates are set.

The resulting information objects can be presented in different map applications.

4. CASE STUDIES

Two case studies were processed to prove the correctness and reliability of the automatic recognition of locations based on the work of Alexander von Humboldt [4] and Johann Wolfgang von Goethe [3].

Humboldt's works are presented in the Humboldt Digital Library [5] and the Humboldt Portal [6] where semi-automatic identification of locations made it possible to present the travel routes (see Figure 3) on the basis of Google Maps [2]. The case study of the same source but with an automatic recognition process detected most of the locations in South America, which was indeed one of the travel regions of Humboldt.

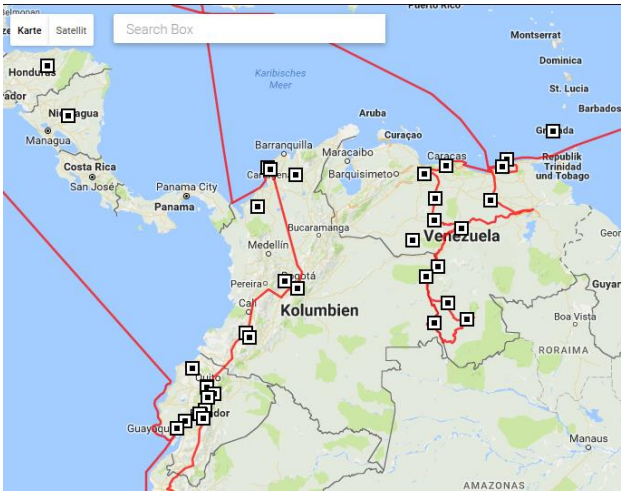

Figure 3: Locations and travel routes of Alexander von Humboldt to the Americas, published as travel literature and natural studies, and diaries, presented in the Humboldt Portal [6].

In comparison to the presentation of Humboldt's travels in the Humboldt Portal, which bases on text analysis and partially on literary work, the automatic recognition of locations generates a remarkable coincidence of the resulting maps (See Figure 4).

Figure 4: Automatic detected locations of the travels of Alexander von Humboldt, representing the objects, which are classified as travel locations.

Another case study was worked out using the "Italienische Reise" of Johann Wolfgang von Goethe [3] because of the well-known travel route and the defined locations (See Figure 5) which Goethe mentioned in his work.

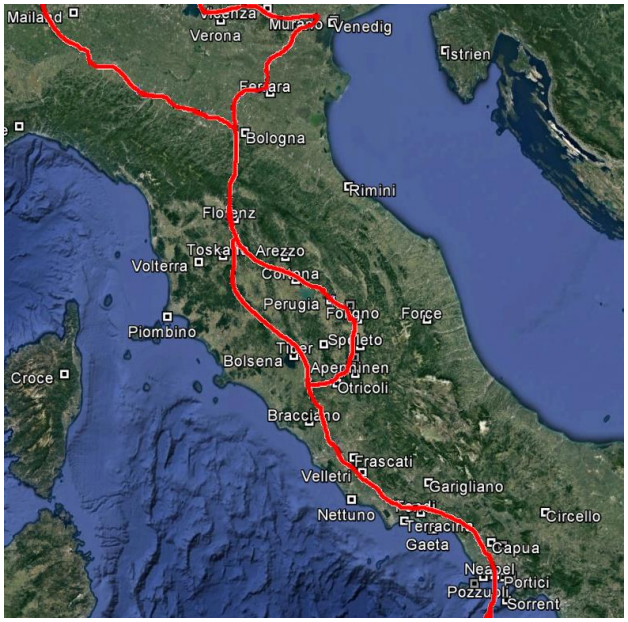

Figure 5: Travel route of Johann Wolfgang von Goethe (Italienische Reise) from Wikipedia as an overlay in Google Earth with the localities coming from the automated process of recognition.

In detail, the results reveal numerous errors due to misspellings, change of location names, equality of terms and location names. On the other hand, it becomes very clear that the automatic object detection and recognition can be improved by error-free and comprehensive sources. As a result, an increase in quality and usability of the service can be expected, accompanied by more options to detect unknown locations in the descriptions of rare books.

5. FURTHER IMPROVEMENTS

Due to redundancies of worldwide location names and to focus on relevant locations, it is necessary to make further improvements. By building a geographical corridor around e.g. the mentioned travel routes or the travel region, the dataset can be reduced by elimination of multiple false positives in recognition of location names. Another option is to analyze regional dependencies on the location points by geostatistical methods. Both processes can be implemented as methods to identify the information object, which is given by the term in the rare book and classified as a geographic location by methods of data mining.

.

6. REFERENCES

[1] K. Coyle: "Mass Digitization of Books"- Journal of Academic Librarianship, v. 32, n. 6, 2006]. [2] D. Doherr & F. Baron, Humboldt: "Digital Library and

Interconnectedness", The Environmentalist, ISSN 0251-1088, DOI 10.1007/s10669-011-9369-y, Springer-Verlag, 2011
[2] Detlev Doherr (2016): „Alexander von Humboldt's idea of interconnectedness and its relationship to interdisciplinarity and communication"- International Multi-Conference on Complexity, Informatics and Cybernetics IMCIC, Orlando, 2016, USA
[3] Johann Wolfgang von Goethe: Italienische Reise.- 1829
https://archive.org/details/goethesitalienis00goetuoft
[4] Alexander von Humboldt, Aimé Bonpland, Hermann Hauff: Reise in die Aequinoctial- Gegenden des neuen Continents: in deutscher Bearbeitung von Hermann Hauff.- Vol 1, Verlag J. C. Cotta, New York, 1862
[5] Humboldt Digital Library - http://avhumboldt.net (Last access: Sept 2016)
[6] Humboldt Portal - http://humboldt.hs-offenburg.de (Last access: Sept 2016)