A brief textual analysis of a corpus on digital libraries and text mining

Mathieu Andro

A brief textual analysis of a corpus on digital libraries and text mining

Mathieu Andro
http://bibliotheque-numerique.fr

Summary

Over 40 million books have been digitized by Google Books. Libraries around the world have participated in this movement to digitize print heritage. The digital corpora now available contain textual data which today is to be extracted using text mining. Specialised in digitization and digital libraries and with experience in text mining, we wanted to make a state of the art on the subject of text mining applied to digital libraries using text mining technologies themselves.
We are indeed wondering about these technologies. Is it developing? Does it work? Are we making discoveries?
We have assembled a corpus of metadata and summaries from Google Scholar which seems to be the most exhaustive source on the subject of scientific but also professional literature. We performed textual analyzes using CorTexT, a tool developed by research and higher education in France

Introduction

Text mining is probably a natural outlet for library digitization projects. We wish to verify that and to analyze studies published on the subject.

Method

1- We asked Google Scholar, which seems to us to be the most exhaustive bibliographic database on the subject. We used the following query:

("digital library" OR "digital libraries") AND ("digitisation" OR "digitization" OR "digitized" OR "digitised") AND ("text mining" OR "Natural language processing")

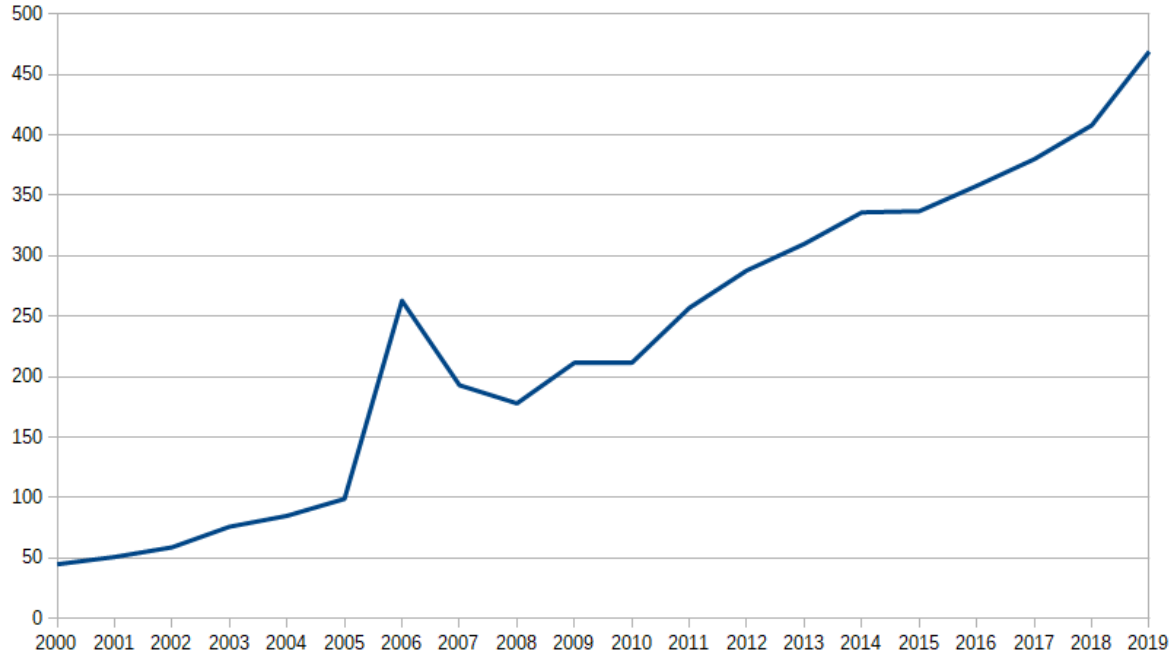
2- We extracted the metadata of 4 830 publications from Google Scholar via Harzing's Publish or Perish

3- We export them in RIS standard.

4- We analyzed the corpus with CorTexT (https://www.cortext.net/projects/cortext-manager).

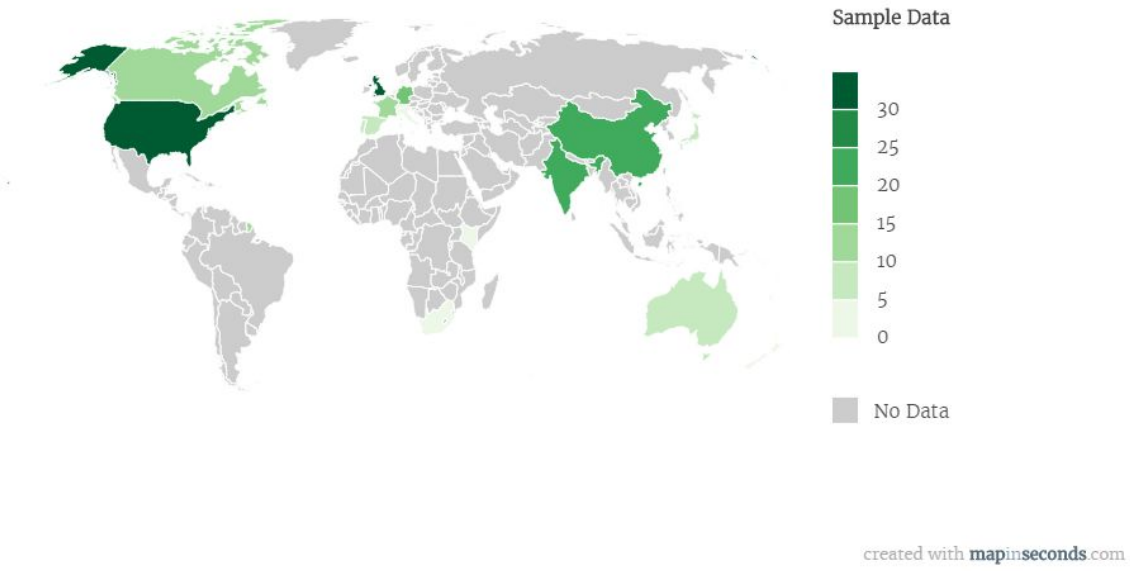
Results

Is it a growing subject ?

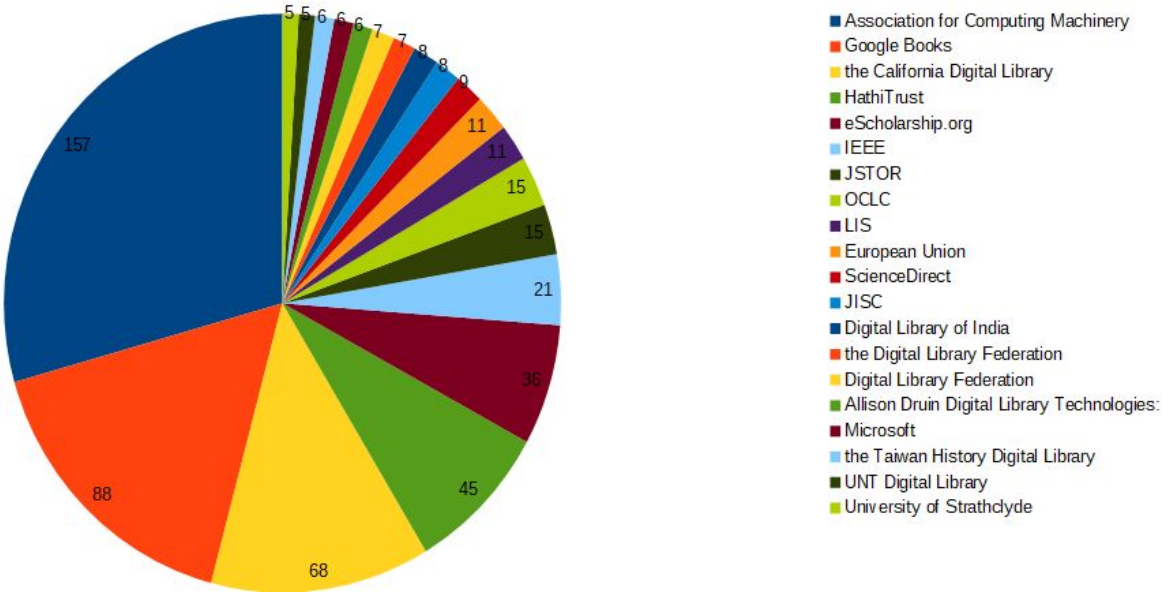


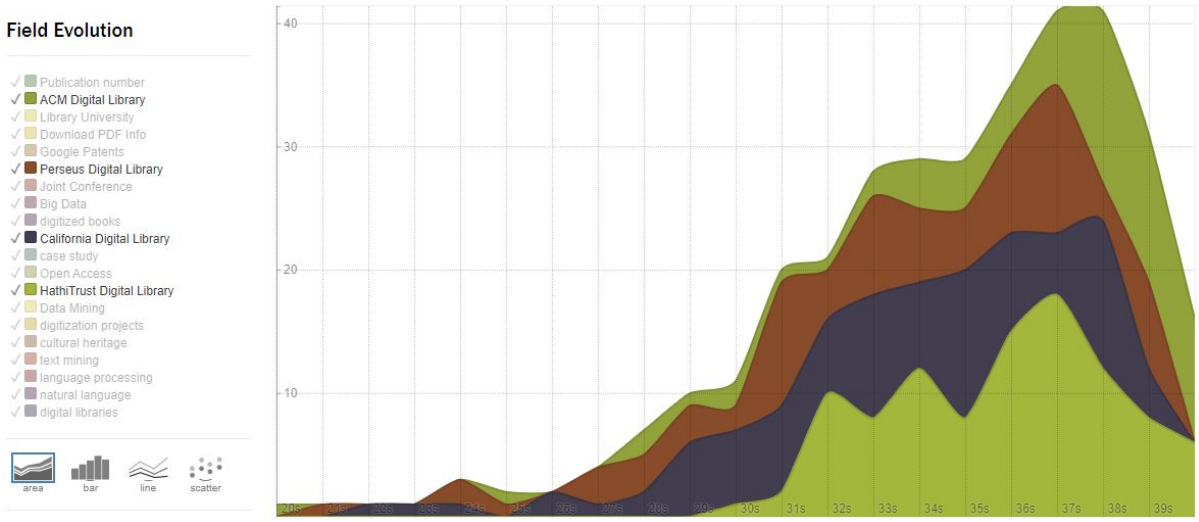
Who works on this subject ?

In which countries ?

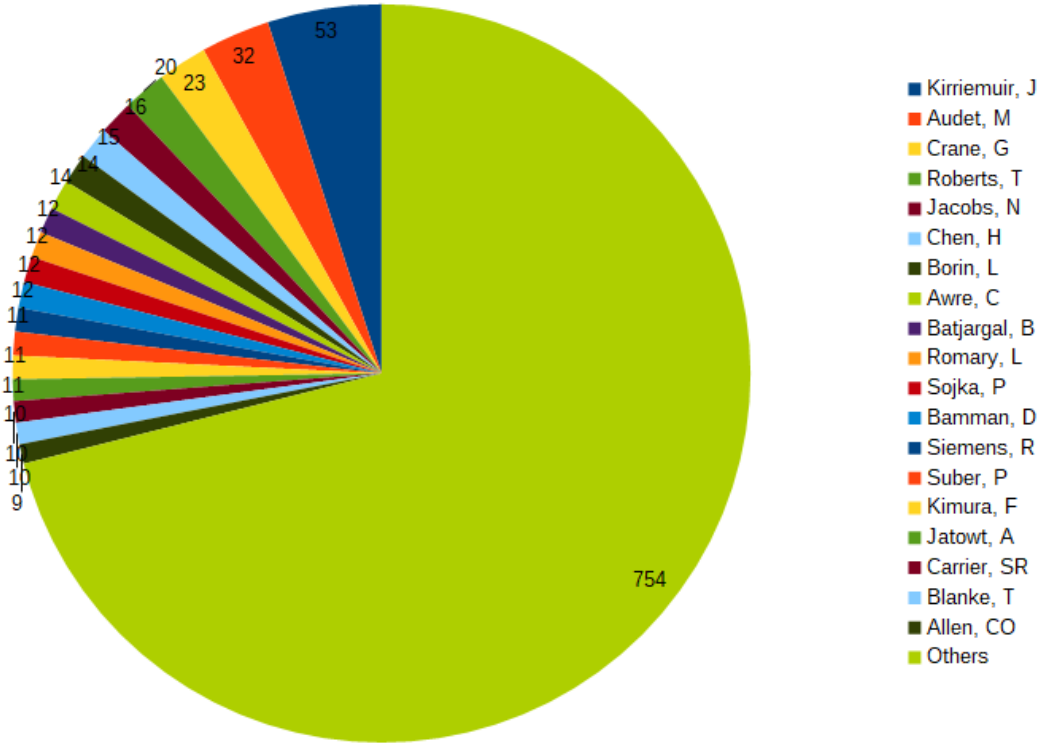


Which institutions and projects ?



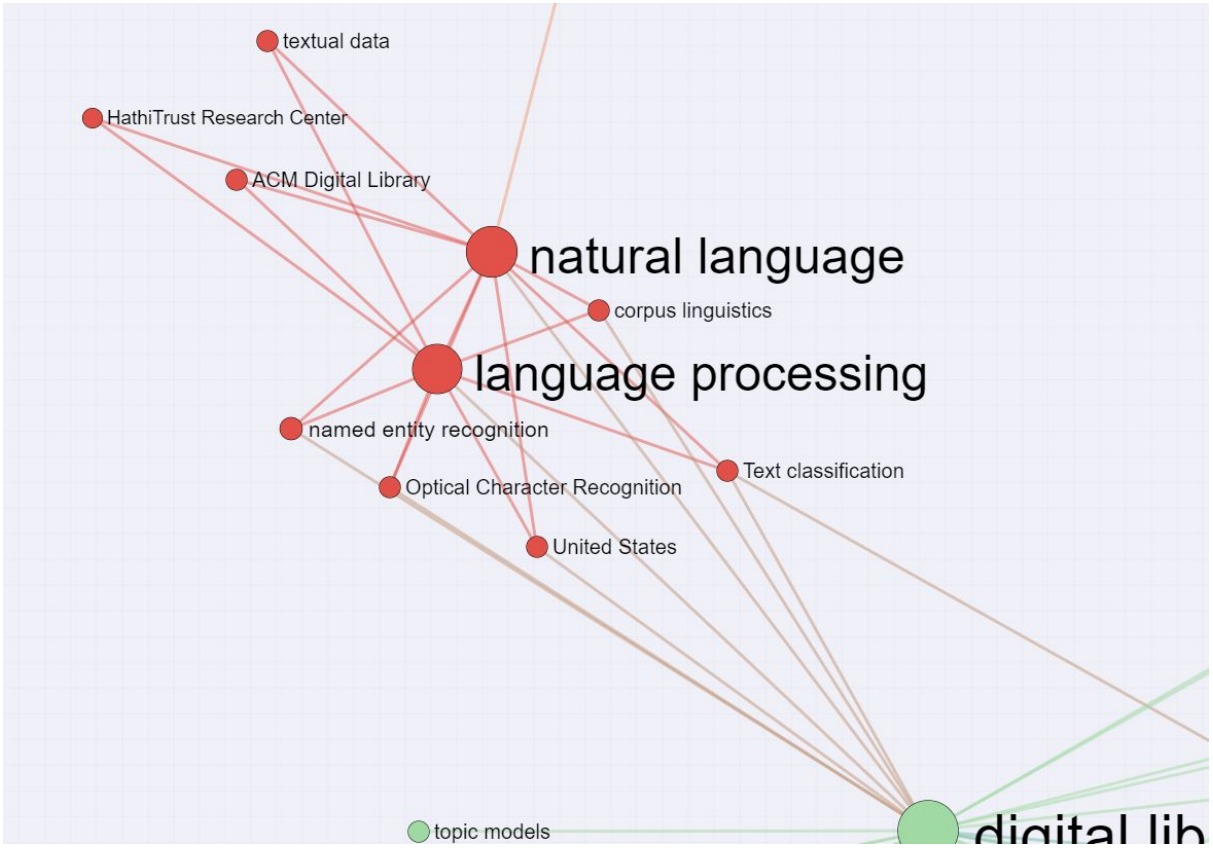


Which authors ?

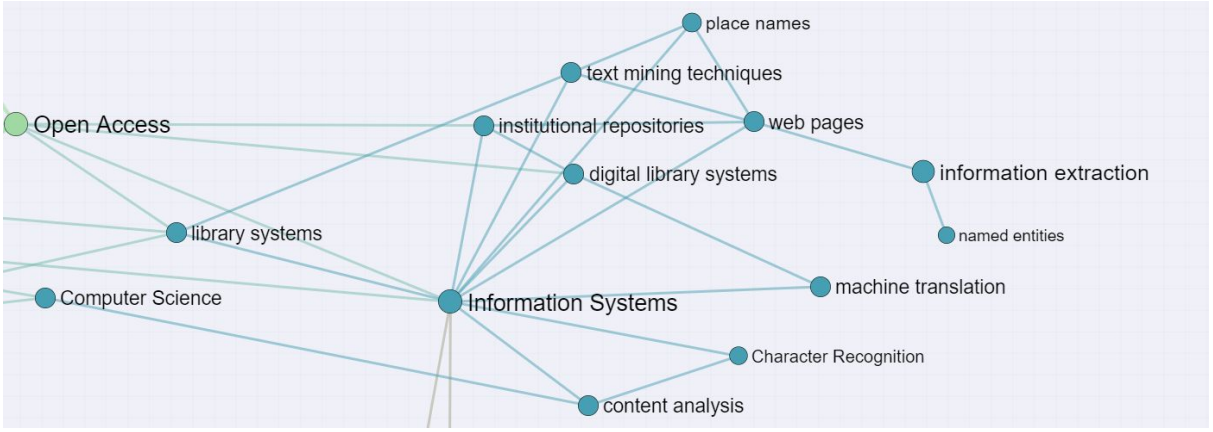


What about are they deal with ?

A central group deals with text mining using digital libraries:
Concepts: open access, artificial intelligence, knowledge management and digital age
Technologies: linked data, entity recognition, improving OCR accuracy, document image analysis
Projects: Perseus Digital Library project

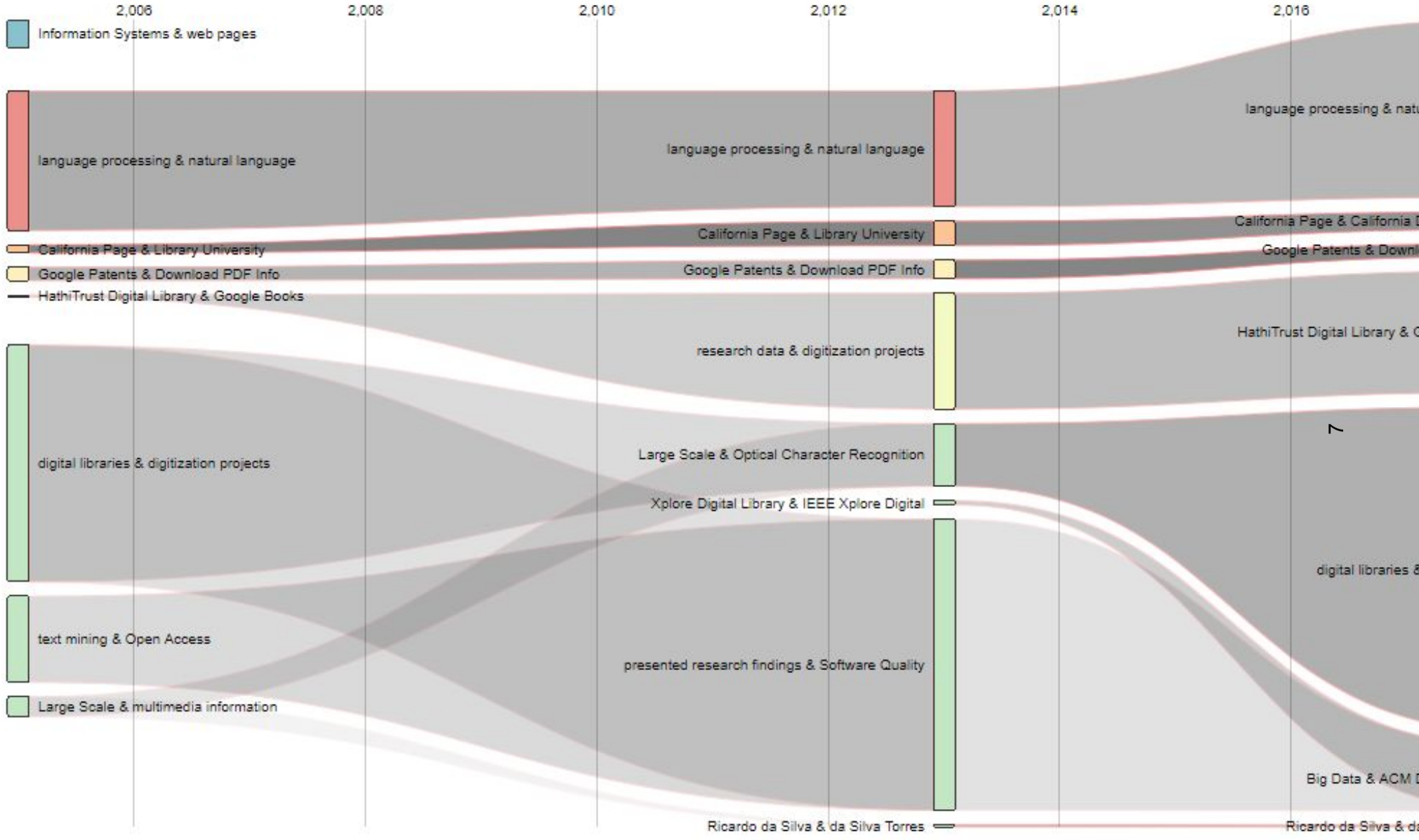


A second group of publications is about natural language using digital libraries:
Concepts: natural language, language processing
Technologies: text classification, Optical Character Recognition (like the first group) and named entity recognition (like the first group)
Projects: Hathi Trust Research Center, ACM Digital Library



A third one is about digital libraries using text mining technologies:
Concepts: Information Systems, Institutional repositories
Technologies: Machine translation, named entities (like the 2 other groups), place names, character Recognition (like the 2 other groups)

Finally we also obtain a group for Hathi Trust and Google Books, and another one for California Digital Library


2,006
2,008
2,010
2,012
2,014
2,016
Information Systems & web pages
language processing & natural language
California Page & Library University
Google Patents & Download PDF Info
HathiTrust Digital Library & Google Books
digital libraries & digitization projects
text mining & Open Access
Large Scale & multimedia information
language processing & natural language
California Page & Library University
Google Patents & Download PDF Info
research data & digitization projects
Large Scale & Optical Character Recognition
Xplore Digital Library & IEEE Xplore Digital
presented research findings & Software Quality
Ricardo da Silva & da Silva Torres
language processing & natu
California Page & California
Google Patents & Down
HathiTrust Digital Library &
digital libraries &
Big Data & ACM
Ricardo da Silva & da

If we have a look at temporal evolution of all theses concepts, it seems that Hathi Trust Text Mining Projects are a growing subject. Large scale is also a growing subject.

References

Witten, I. H., Don, K. J.,Dewsnip, M., Tablan, V. (2004). Textmining in a digital library, International Journal on Digital Libraries, 4: 56–59

Sanderson, Rob & Watry, Paul. (2007). Integrating Data and Text Mining Processes for Digital Library Applications. 73-79.

Robert B. Allen (2010). Improving Access to Digitized Historical Newspapers with Text Mining, Coordinated Models, and Formative User Interface Design. IFLA Newspaper Section Meeting, New Delhi, February 2010

Fox, R. (2010). Mining the digital library. OCLC Systems & Services: International digital library perspectives, Vol. 26 No. 4, pp. 232-238.

Dietmar Wolfram (2016). Bibliometrics, Information Retrieval and Natural Language Processing: Natural Synergies to Support Digital Library Research. BIRNDL 2016 Joint Workshop on Bibliometric-enhanced Information Retrieval and NLP for Digital Libraries

Philipp Mayr, Muthu Kumar Chandrasekaran, Kokil Jaidka (2017). Report on the 2nd Joint Workshop on Bibliometric-enhanced Information Retrieval and Natural Language Processing for Digital Libraries (BIRNDL 2017). ACM SIGIR Forum 107 Vol. 51 No. 3 December 2017

Philipp Mayr, Muthu Kumar Chandrasekaran, Kokil Jaidka (2018). Report on the 3rd Joint Workshop on Bibliometric-enhanced Information Retrieval and Natural Language Processing for ACM SIGIR Forum 105 Vol. 52 No. 2 December 2018

Gesare Asnath Tinega, Waweru Mwangi, Richard Rimiru (2018). Text Mining in Digital Libraries using OKAPI BM25 Model. International Journal of Computer Applications Technology and Research, Volume 7–Issue 10, 398-406.

Philipp Mayr, Ingo Frommholz, Guillaume Cabanac, Muthu Kumar Chandrasekaran, Kokil Jaidka, Min-Yen Kan, Dietmar Wolfram (2018). Introduction to the special issue on bibliometric-enhanced information retrieval and natural language processing for digital libraries (BIRNDL). Int J Digit Libr 19:107–111

Dickson Koehl, E., Dubnicek, R. (2019). Text Mining with HathiTrust. 2019 ACM/IEEE Joint Conference on Digital Libraries