

EGOLink: Supporting Editors of Online Historical Sources through Automatic Link Discovery

Hatem Mousselly Sergieh, Michael Piotrowski, Iryna Gurevych

1. Introduction

EGO (European History Online) is a transcultural history of Europe on the Internet published by the Leibniz Institute of European History. The success and the consequential growth of EGO is, however, a challenge for the editorial office. The interlinking of EGO articles with each other and with external resources is an important aspect of EGO's conceptual design, so each new article has to be integrated into the existing link structure. This means not only that the new article has to be linked to relevant existing articles or online sources, but the copy editors must also check whether links to the new article need to be added to existing articles.

Doing the linking manually is a tedious task. Therefore, we aim in the EGOLink project at developing methods for (semi-)automatically linking EGO articles in order to reduce the manual effort for the editorial office and to improve the navigation for readers.

As a first step, we created a tool for visualizing and analyzing the current link structure in EGO document collection. Besides detecting problems such as under-linked articles, the analysis tool provides information that is needed by a (semi-)automatic linking system. In the next step, the automatic generation of links is addressed; the two main research questions here are: (1) the automatic identification and ranking of potential link anchors in EGO articles, and (2) the discovery of suitable link targets in other EGO articles or external resources.

2. Background

2.1. EGO Document Collection

EGO provides a platform for publishing refereed academic articles about the history of Europe with a special focus on modern Europe from the end of the Middle Ages up to contemporary history. EGO is bilingual, i.e., it contains articles in English and German. Most articles written in one language are translated into the other language and published in both languages. Currently, EGO contains 670 articles: 436 in German and 234 in English.

Linking articles to each other as well as to external resources is a major focus of EGO. Once a new article is submitted the editorial staff put a special effort on linking the articles to already existing EGO articles as well as to other multi-media contents, such as images, interactive maps, video and audio clips.

2.2. Entity Linking

As already mentioned, linking articles is at the heart of EGO. So far, this task has been done manually. We aim to optimize this process by developing a (semi-)automatic linking tool based on state-of-the-art approaches while taking the peculiarities of EGO into account.

The problem of (semi-)automatically linking EGO articles to each other or to external resources can be modeled as link discovery, or entity linking (Larson, 2010). In general, entity linking consists of two main steps (Erbs, et al., 2011): 1) anchor discovery, which identifies text mentions that can function as link anchors, and 2) target discovery which identifies the best matching references for each anchor from a set of candidates. While anchor discovery can be handled independently of the document collection at hand – e.g. using NER (Named Entity Recognition) (Nadeau, et al., 2007) or keyphrase extraction methods (Turney, 2000), discovering and ranking the targets are much affected by the nature of the document collection.

Currently, several state-of-the-art algorithms for entity linking have been made available (Hoffart, et al., 2011) (Ceccarelli, et al., 2013). However, applying such algorithms directly to perform entity linking based on EGO is challenging. In general, current algorithms rely on extracting statistics about candidate

entities, e.g., popularity-based priors; form a huge corpus like Wikipedia. The priors are then used during the process of the target discovery and ranking. However, such an approach is not applicable in our case due to the tiny size of EGO compared to other similar resources like Wikipedia. Another challenge is posed by the German part of EGO. Current approaches focus on English and a special effort should be made to address entity linking for German.

(Erbs, et al., 2011) demonstrated that approaches for target ranking that leverage the linking structure available in the document collection perform best compared to other methods that only leverage the titles of the documents or the associated texts. In light of this result, we propose a tool for analyzing the linking structure in EGO's document collection. The proposed tool helps us to understand the current status of EGO's links and to gain insight regarding the decisions that should be taken to develop the actual entity linking system. Besides identifying links among the articles, the proposed tool also extracts information from an additional resource, namely the literature referenced by the articles. Our hypothesis is that articles sharing the same reference should be linked to each other. As a result, this kind of analysis can help the (semi-) automatic entity linking system to identify further anchors, thus the interlinking density of the document collection can be increased.

3. EGO Analyzer

EGO Analyzer is a web-based analysis tool for EGO document collection¹. The tool can handle the English as well as the German part of EGO jointly or separately. Besides providing different kinds of metadata about EGO's articles, the tool allows analyzing EGO from different perspectives: the corpus, the article and the reference perspectives (Figure 1).

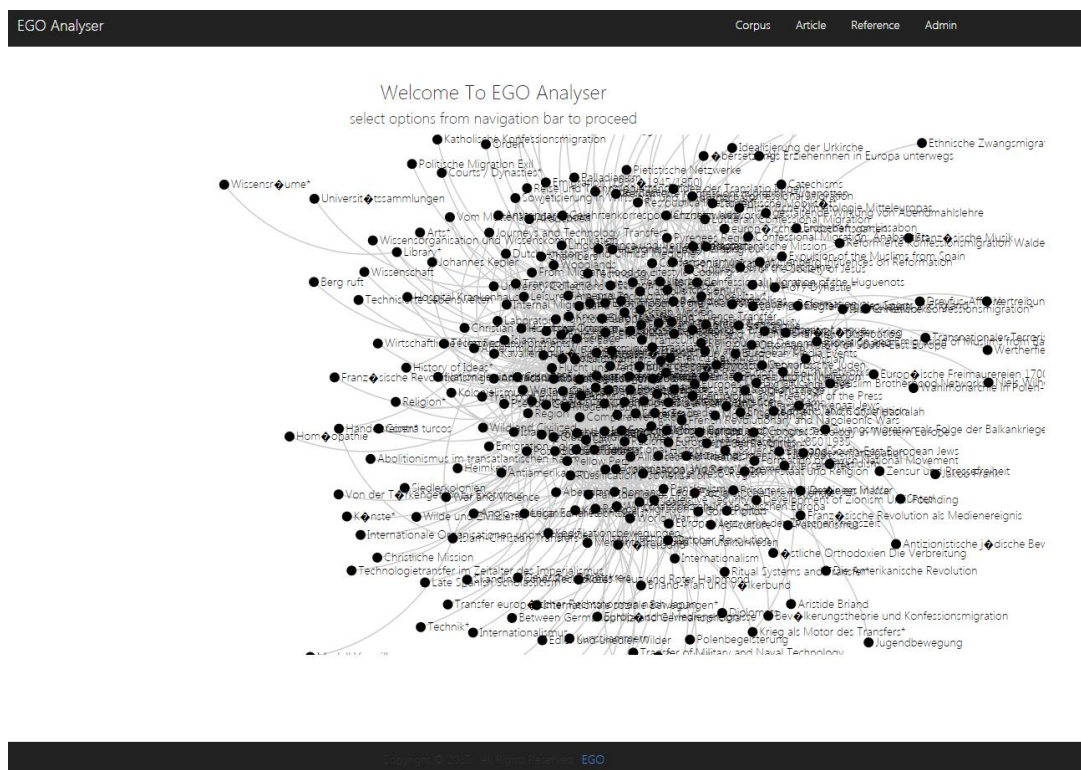


Figure 1 Screenshot: EGO Analyzer

¹ The tool is accessible online at <http://ieg-42.zdv.uni-mainz.de/>

For instance, given the article titled “Science”² from the English collection, the tool extracts links to other articles in both directions, i.e., outgoing as well as incoming links (Figure 2). Furthermore, we can view the set of articles that reference the same literature as the target article (Table 1).

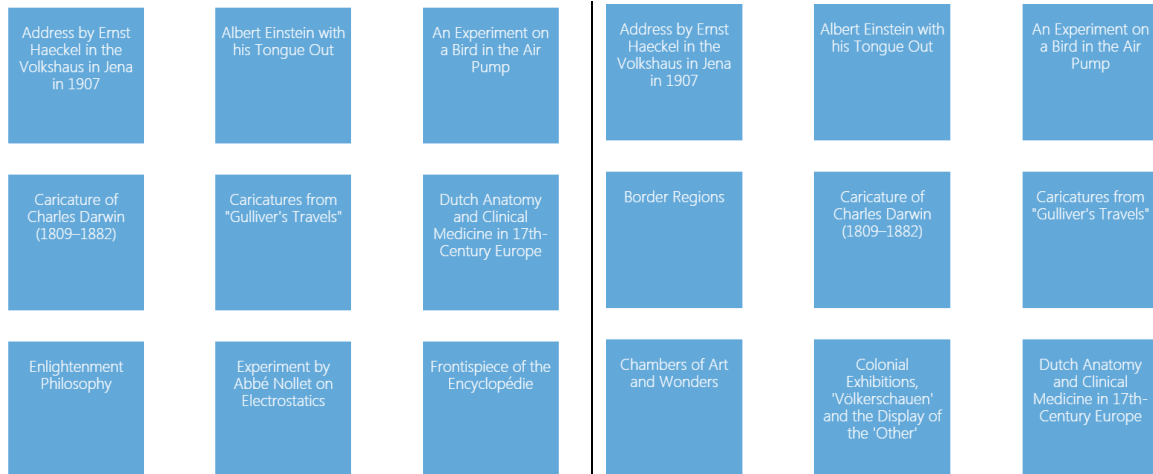


Figure 2: Source article: “Science”

Left: articles referred to by the source article (outgoing links)

Right: articles referring to the source article (incoming links)

Article	#Common References
Wissenschaft	152
Europäische Freimaurereien 1850 1935	33
Der Genfer Psalter	32
Nationale und transnationale Nachrichtenkommunikation	25
National and Transnational News Distribution	25

Table 1: Articles sharing a subset of references with the target article “Science” and statistics of common references

4. Results of the Analysis

We used the EGO Analyzer to derive statistics about the linking structure of EGO. First, we collected statistics about the links available in the document collection. For this purpose, we divided the links into two categories: internal and external. Internal links are those which point from one EGO article to another, while external links refer to external targets, such as entries in the Virtual International Authority File (VIAF). For each article, we also distinguish between two types of links: incoming and outgoing links. Figure 3 shows the average number of internal/external as well as the average number of incoming/outgoing links per article. The results show the dominance of the external links (3 times as many as internal links). This can be explained by the relatively small size of the document collection which makes it hard for the editorial staff to find corresponding internal links for entities of a newly added article. This also explains the higher number of outgoing links per article. In fact, most of these links go to external resources which, in turn, are not necessarily linked back to EGO articles.

This observation demonstrates the urgent need for an automatic link discovery system that enables the editors to enhance the interconnectivity among EGO articles. The same system can also run as a background process that checks for candidate links each time a new article is added.

² <http://ieg-ego.eu/en/threads/backgrounds/science>

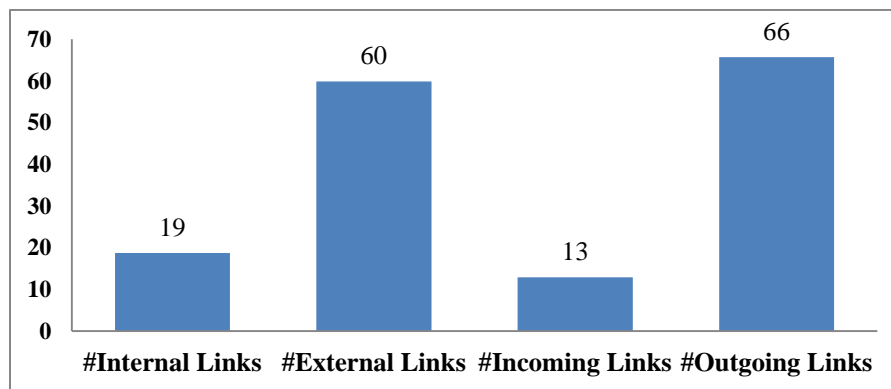


Figure 3: Link analysis of the EGO document collection

5. Conclusion

In this abstract, we presented work in progress for assisting editors of EGO to (semi-) automatically link the articles. We focused on the first step towards such a system, namely, analyzing the link structure in the target document collection. We presented a web-based tool that is able to perform such analysis. Currently, we are working on the actual link discovery system as well as a user-friendly interface for the editorial office. By the time of the conference, we will also provide a demonstration of the actual linking tool.

References

- Erbs, Nicolai , Zesch, Torsten and Gurevych, Iryna. 2011.** Link discovery: A comprehensive analysis. *Semantic Computing (ICSC), 2011 Fifth IEEE International Conference on.* 2011, pp. 83--86.
- Ceccarelli, Diego, et al. 2013.** Dexter: an open source framework for entity linking. *Proceedings of the sixth international workshop on Exploiting semantic annotations in information retrieval.* 2013, pp. 17--20.
- Hoffart, Johannes, et al. 2011.** Robust disambiguation of named entities in text. *Proceedings of the Conference on Empirical Methods in Natural Language Processing.* 2011, pp. 782--792.
- Larson, Ray R. 2010.** *Information retrieval: Searching in the 21st century; human information retrieval.* s.l. : Wiley Online Library, 2010.
- Nadeau, David and Sekine, Satoshi. 2007.** A survey of named entity recognition and classification. 2007, Vol. 30, 1, pp. 3--26.
- Turney, Peter D. 2000.** Learning algorithms for keyphrase extraction. 2000, Vol. 2, pp. 303--336.