A Knowledge Graph from the *Regesta Imperii*: Construction, Visualization and Macro-level Analyses

Leo Born, **Juri Opitz**, and **Vivi Nastase**
Department of Computational Linguistics
Heidelberg University
69120 Heidelberg, Germany
{born,opitz,nastase}@cl.uni-heidelberg.de

Motivation

Being able to visualize and analyze large-scale textual data has multiple benefits: it allows for the inspection of large amounts of data; finding new and interesting regularities and patterns that are apparent only through a higher-level view of the data; and testing theoretical hypotheses about larger-scale or distant connections that are more difficult to achieve through purely manual analysis.

We built a knowledge graph from the *Regesta Imperii* data set that formalizes multiple layers of information from the texts, and we provide an easy-to-use exploration tool for this graph. Our tool is a web application that opens the access to this repository of medieval documents and allows for various types of analyses that we believe would be of interest to historians and other humanities researchers. The data and application are shared publicly.[1]

Regesta Imperii – Abstracts of Medieval Charters

The *Regesta Imperii* (RI) is a large-scale corpus for medieval European history studies. It consists of documents, *Regests*, that can be seen as abstracts of charters issued

[1] https://gitlab.cl.uni-heidelberg.de/born/ri-visualization.

by Roman-German emperors and popes, starting from the Carolingian dynasty (8th century) to Maximilan I. (16th century). More than 175,000 Regests have been converted to Unicode and are stored in a publicly available online database.[2]

So far, not much computational research has been conducted on the RI. Kuczera (2015) [1] projected attributes and relations between entities from the times of Friedrich III. (i.e. a subset of the RI) into a graph database, relying on a manually created person register for the universe of Friedrich III. Opitz and Frank (2016) [2] manually labeled 500 randomly sampled Regests with 12 medieval themes and players of interest (e.g. *nobles*, *spiritual institutions*, or *war and peace*) and trained binary classifiers to label all Regests and compute statistics about the importance of the medieval themes and players with regard to time.

Inducing a Knowledge Graph from the RI

Many Regests consist of only one sentence, describing an action performed by the issuer (usually an emperor or pope) towards one or several of his subjects. We induced a knowledge graph (KG) of the *Regesta Imperii* by automatically extracting such relation triples using Natural Language Processing (NLP) methods. We rely on the fact that the main relation (edge) between a ruler and its subject entities (the nodes) occurs very frequently in a subject-verb-object format. Because the main sentence contains this relation information, we process only this sentence and perform syntactic parsing and named entity recognition on it to extract the main verb (relation) and the subject entities; the issuer of a relation is extracted from the Regest's metadata. We bootstrap our preprocessing and relation extraction using information particular to the dataset: we noticed that a sentence split was often performed on nobility titles prefixing named entities: e.g. *Gf.* – Graf (count) or *Bf.* – Bischof (bishop). By filtering out frequent titles appearing in front of named entities, we reduced erroneous sentence splits by 10%. The titles with their typical placement in front of named entities further allowed us to enhance the KG structure with additional nobility attributes for nodes. For more details, we refer the reader to [3].

Our induced, directed multi-edge knowledge graph of the *Regesta Imperii* contains 68,574 nodes and 154,097 edges. The average outdegree of nodes is 432 and average indegree is 2.25, showing that the KG is highly uni-directional with few central nodes (i.e. issuers) and most peripheral nodes having few incoming and almost no outgoing edges. Structurally speaking, the knowledge graph is weakly connected, consisting

[2] http://www.regesta-imperii.de/en/home.html.

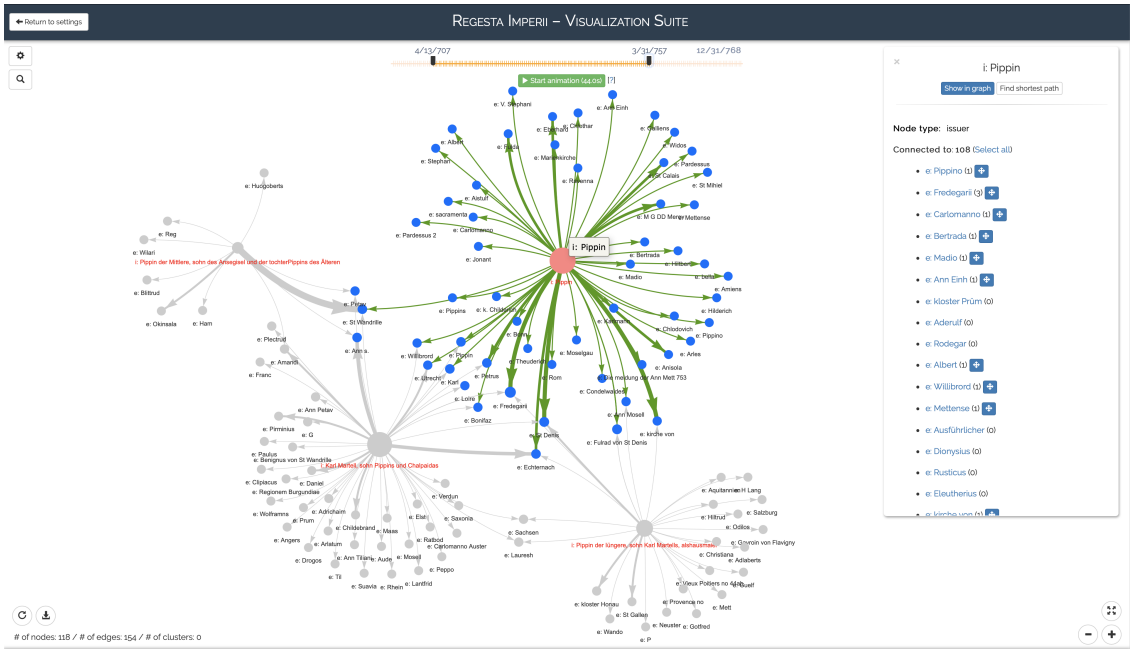

Figure 1: Example graph of the four issuers king *Pippin* (highlighted), *Pippin der Iüngere* (before he was king), *Karl Martell* (father of Pippin) and *Pippin der Mittlere* (predecessor of Pippin) between 4/13/707 and 3/31/757 (118/180 nodes, 154/245 edges). Edge thickness indicates edge weight and nodes are scaled by degree centrality. The figure is best seen in color.

of 14 components with one large component containing 68,540 and the remaining ones containing < 5 nodes. Analysis of nodes based on centrality betweenness and the set of cut vertices[3] showed that most of these nodes are persons, and roughly 12% denote places or regions, indicating that certain regions also play a critical role in connecting multiple sub-networks.

Visualizing the Knowledge Graph

Visualizing an extremely large graph such as the one we built from the RI has no analytical benefits – higher-level patterns or detailed information would be obscured by the sheer size of the graph. We use optimizations and ad-hoc querying to visualize

[3] These are the nodes whose removal would increase the number of components, thus they are nodes critical for network connectivity.

sub-networks of the KG. These include setting a threshold X ($\leq X$; $= X$; $\geq X$) for the number of connections to display for a named entity and the number of relation instances. With these settings, specific requirements can be made, e.g. "show the network of Friedrich III. with relations that appear at least 40 times" and a combination of constraints is also possible. In practice, we are able to create interactive visualizations for networks of up to 12,000 nodes. A small example graph is shown in Figure 1.

Visualization is done using *vis.js*,[4] and the resulting networks are interactive, meaning that elements can be searched, highlighted, or hidden, either manually or automatically based on various filters, including a temporal filter (*time slider*) and an attributional one (e.g. based on nobility or named entity type). On top of that, the visualization tool allows basic graph-theoretical calculations and operations on the resulting networks, for example measuring centrality[5] or finding shortest paths or clusters.

Potential Analyses

Our tool allows to select as many issuers for which to build networks as possible, serving different purposes depending on the particular number of issuers selected. Selecting only one issuer is useful when specifically analyzing patterns of interaction of only that issuer. This is what has been done in most previous work on the RI in traditional humanities research (e.g. [4, 5]). We believe, however, that having a digital resource such as our tool at hand is also beneficial for this use case.

The greatest benefit our tool provides is more high-level macroanalysis. For example, selecting multiple issuers allows to comparatively analyze interactional patterns of dynasties, "similar" or "dissimilar" issuers, or any arbitrary issuers. "Similarity" can be read in an attributional sense, e.g. two kings who were said to be benevolent. This would allow for a comparison of their interactional patterns, i.e. "What kind of relations were dominating the Regests of these kings and do they align with their image of benevolence?". Furthermore, *structural* similarity, e.g. in the form of network assortativity,[6] can also be considered because mixing patterns of the networks might yield insights into issuer networks as well.

Furthermore, we include functions to cluster nodes based on additional properties in the form of nobility titles or named entity types. This enables a more generalized

[4] `visjs.org`. This is an open-source tool aimed specifically at network visualizations.
[5] Currently, we support degree and eigenvector centrality.
[6] This describes the tendency with which nodes are associated with similar nodes. See Newman (2010, pp. 220-231)[6] for more details on the types of similarity and their calculations.

analysis of the data by means of abstract concepts (e.g. "How many interactions with abbots did Charlemagne have between 800 and 814?" – 19). Similarly, we can make use of the fact that there are city names that have the German term for city – *Stadt* – in them, in order to simply cluster entities based on surface form.

A different form of enrichment comes in the form of edge attributes. While preprocessing the RI, we associated an attribute list to each edge containing date and location of a Regest, as well as key phrases associated with each event relation.[7] By formulating a text classification task on the main relation of each Regest using all noun chunks and verbs as features, we ranked the phrases describing an instantiation of a relation according to the learned weights for the relation. This allows for higher-level analysis of the Regests by means of content as well, allowing specifically to highlight relations based on key phrases. Using such a key phrase filter for example shows that associated with Maximilian I. are many financial relations such as *schuldet* (owes), *verbucht* (books) and *bezahlt* (pays) – Maximilian I. was well known for his debts he accumulated due to many wars and a rather extravagant lifestyle and this is also statistically reflected in the RI.

Lastly, we want to give researchers the ability to query the data in accordance with their own research questions. We also envision employing the web application as an educational tool, by providing it to easily query, visualize, and contextualize key medieval players and their networks. This would facilitate the students' ability to comprehend and analyze historical events from a prosopographical perspective by employing a *distant reading* [7] methodology. Doing so would not only further the intersection of humanities education and digital tools and resources, it would also provide a common platform for enabling prospective scholars to internalize and employ a network-centric methodology for history studies.

References

[1] A. Kuczera, "Graphdatenbanken für Historiker. Netzwerke in den Registern der Regesten Kaiser Friedrichs III. mit neo4j und Gephi," *Mittelalter. Interdisziplinäre Forschung und Rezeptionsgeschichte*, 2015.

[2] J. Opitz and A. Frank, "Deriving Players & Themes in the Regesta Imperii using SVMs and Neural Networks," in *Proceedings of the 10th SIGHUM Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities*, pp. 74–83, 2016.

[7] For example, at one time king Sigmund may have promised *bestowal of land* to duke Ludwig and at another time he might have promised him *privileges* or *financial help*.

[3] J. Opitz, L. Born, and V. Nastase, "Induction of a Large-Scale Knowledge Graph from the Regesta Imperii," in *Proceedings of the 2nd Joint SIGHUM Workshop on Computational Linguistics for Cultural Heritage, Social Sciences, Humanities and Literature (LaTeCH-CLfL)*, (Santa Fe, New Mexico, USA), pp. 159–168, 2018.

[4] D. Bulach, "Organisieren von Herrschaft im späten Mittelalter. Ludwig der Bayer und der Nordosten des Reiches," in *Ludwig der Bayer (1314 - 1347). Reich und Herrschaft im Wandel* (H. Seibert, ed.), pp. 263–284, 2014.

[5] J. Laczny, "Friedrich III. (1440-1493) auf Reisen. Die Erstellung des Itinerars eines spätmittelalterlichen Herrschers unter Anwendung eines Historical Geographic Information System (Historical GIS)," in *Perzeption und Rezeption. Wahrnehmung und Deutung im Mittelalter und in der Moderne* (J. Sarnowsky, ed.), pp. 33–65, 2014.

[6] M. Newman, *Networks. An Introduction.* Oxford, United Kingdom: Oxford University Press, 2010.

[7] F. Moretti, *Distant Reading.* London: Verso, 2013.