The potential of LLMs for constructing a socio-legal knowledge graph

Christian Boulanger

1. Knowledge graphs as a tool to study disciplinary knowledge production

Why and how do academic disciplines and research fields emerge and evolve? This is one question that the field of the history and sociology of science is concerned with (Stichweh, 1992). When one surveys the literature, it becomes apparent that traditionally, a major focus has been on the natural sciences, and it is only recently that the social sciences or humanities are receiving more attention (Kuznetsov, 2019). I am looking into the question of disciplinary development in the context of my research on the post-war history of the German-language sociology of law, which was unable to institutionalize itself as an academic (sub-)field in a way similar to the U.S. or the UK, despite Germany's rich pre-1945 socio-legal intellectual history (Machura, 2020).

The traditional way to study the history of an academic discipline has been the "close reading" of the available literature, the study of the biographies of scientists and scholars, archival work on the actors and the institutions that have been crucial in the development of the discipline. Hermeneutic methods remain essential in grasping the complex history of a research field. But in recent years, the use of computational methods has been on the rise not only being increasingly used in the fields of the digital humanities (Luhmann and Burghardt, 2022), digital history (Cohen et al., 2008), or digital legal research (Livermore, 2019), but also in the field of the history of science (Gibson and Ermus, 2019). Digital Humanities methods thus provide an additional perspective on disciplinary history. Their promise is to discover and reliably document structures and relationships that only appear when algorithmically analyzing a large body of data (Jänicke et al., 2015), often referred to as "distant reading" (Underwood, 2017). They allow us to model and visualize these structures and develop metrics for numerical results and comparisons that can be reproduced by others. This is in contrast to hermeneutic studies where we need to trust the plausibility of the author's interpretation and have no access to the body of data on which she based this interpretation on, beyond those sources that are quoted in the scholarly work.

To be sure, this does not mean that computational methods by definition provide more "objective" results. As with any quantitative studies, the selection of the data that is included involves (unacknowledged) value judgements. This is not only limited to data collection and selection. It also affects preprocessing, processing, and, in particular, visualization. As Gibson and Emus express it, rather than bringing research closer to objectivity, "each step of the workflow adds another layer of subjectivity" (Gibson and Ermus, 2019: 555). This has to be kept in mind when interpreting the results of our computational methods.

For the purpose of a Digital Humanities approach, I understand the history of a research field as the history of knowledge production in this field. In contrast to an understanding of scientific knowledge as the process of progressively eliminating false believes by empirical evidence or logical reasoning and replacing them by (temporary) "true" beliefs, the focus of knowledge production centers on the social processes involving human actors in tempo-spacial contexts that influences the form and content of the knowledge they produce in textual form. These processes can be modelled by knowledge graphs, which store information on actors, their products, and the context as a network of entities and the relationships between them (Haslhofer et al., 2018). To be sure, this is an incomplete picture of knowledge production. These methods cannot tell us anything about the context in which these publications have been produced, nor can they capture the real-life processes of knowledge production. Any insights to be gained by taking a bird's-eye view always have to be checked against studies that rely on "close reading".

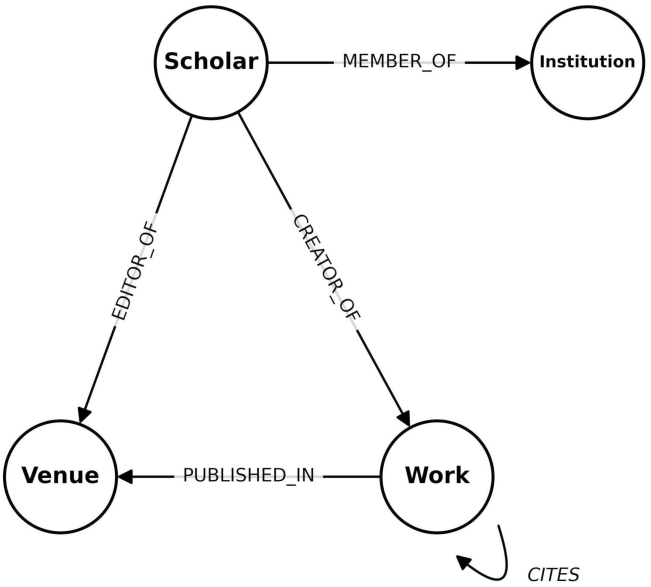
A knowledge graph always needs a data model which describes what entities and relationships the model contains. This data model is an extremely reduced representation of the empirical reality which gave rise to the research question, and, at a minimum, the model must contain the entities that are necessary to answer it. However, not all the entities or relationships that one would like to include in the model can be populated with existing data easily. There is a tradeoff between the complexity of the model and the costs (in terms of time and money) necessary for data acquisition. In my case, the most interesting data points are scholars, works, venues and institutions.

Fig. 1 presents a possible data model for the graph. The entity types "scholars" and "institutions" are self-explanatory. "Works" refer to scholarly publications such as articles, chapters, or books; "venues" refer to containers in which such works are included or of which they are a part, such as academic journals, edited volumes, or book series. The relationship "EDITOR_OF" can express the editorship of books, but an additional purpose is to represent the involvement in the production of serial publications which are represented by the "venue" entity type.

Finally, works can cite other works. This relationship looks inconspicuous in the data model, but is this edge in the graph which contains most of the information. A distinctive characteristic of scholarly knowledge production is the use of references and citations, i.e. the rigorous statement of the literature used, sometimes with the information on the exact location within this literature where an idea or quote has been taken from. In the data model, this information is represented by the CITES relationship between Work nodes. In fact, this is one of the most interesting parts of our data model in terms of the research question. If we know which work cited which other work, it is possible to reconstruct relationships that exist between these works. Needless to say, the

fact that a citation exists does not tell us – by itself – the exact nature of this relationship, let alone answer the question whether one work was influenced by another. There are certain questions that could only be answered by the inclusion, into the data, of the citation context/intent, which tell us, for example, if a reference was cited approvingly or in opposition, for what purpose it was cited, how often or at which particular page (On this, see Small, 2011; Berrebbi et al., 2022. See also Liesegang and Gläser (2026) , Simons (2026), and Simons et al. (2026). For the purpose of the current study I am interested only in the fact whether a reference can be proven to exist.

Fig. 1: The data model of the (socio-)legal theory knowledge graph



What is missing in the visualization of the data model is the temporal aspect. For example, scholars belong to one or more institutions for a specific time period. Thus, for example, the "MEMBER_OF" relationship between "scholar" and "institution" nodes should have a temporal property that contains the information from which start date to which end date the affiliation existed. While this is completely intuitive for humans, the modelling, representation and analysis of temporal relationships is computationally not straightforward and the subject of research on temporal knowledge graphs (Gottschalk and Demidova, 2018; Zhang et al., 2021).[1]

1 There are many more data points that could be added to this model. For example, it would be fascinating to include information on academic conferences which scholars attended. Unfortunately, no databases exist in which information on academic conferences could be simply retrieved from. Instead, getting this data would require text-mining of conference web pages or printed conference proceedings.

2. Challenges for populating the knowledge graph: lack of data and tools

Although this seems like a relatively simple model, populating the model with data encounters many non-trivial challenges. Theoretically, a data source exists for each of the data points in the model. For example, the Research Organization Registry (ROR) contains metadata on institutions (https://ror.org/). Metadata on German Scholars can be retrieved from places like the German Integrated Authority File (GND) (https://www.dnb.de) through a linked data interface (https://lobid.org), which also gives access to bibliographic data on published books. For recent publications (roughly the last two decades), data on journal articles can be found in bibliometric databases such as the Web of Science, Scopus, or OpenAlex.

In the context of a knowledge graph, one major challenge is to disambiguate and link datasource entries reliably. How do we know, for example, if a certain "John Smith" has published an article, which of the many persons of that name is the given author? Specialized databases such as orcid.org which provide unique ids are of very recent origin and only cover a miniscule subset of authors.[2]

The disambiguation and linking problem becomes only relevant once one has large datasets exported from unrelated data sources which need to be consolidated in the knowledge graph. The main challenge the current project faces, in contrast, is the *lack* of data for this particular domain of knowledge. Bibliographic databases such as the one mentioned mainly contain data on contemporary, English-language literature from the life sciences and biomedicine, the physical sciences, business and technology. In contrast, the coverage for non-English language Humanities, law and social sciences has been quite limited (Hammarfelt, 2016; Gläser and Oltersdorf, 2019; Vera-Baceta, Thelwall and Kousha, 2019; Boulanger, Fejzo and Rimmert, 2025). In addition, the time period covered by the research concerns decades of knowledge production prior to the digital age or the assignment of DOIs. This means that metadata coverage of older literature from the target domain is usually very poor. This is confirmed by testing the mere availability of citation data of the Zeitschrift für Rechtssoziologie in the traditional bibliometric data sources. Fig. 2 shows almost no citation data before the turn of the century, and very little afterwards, as compared, for example, to the Journal of Law and Society (UK).

There is another probable reason for the lack of bibliographic data on German-language law, humanities and social sciences in bibliometric databases: a large part of the journals in this area use footnotes (or endnotes) for references. We don't know the technology that the commercial database vendors use for extracting citations from the journals they index. However, existing open source reference extraction software might offer another clue to why there is such little data. The existing software works well for extracting references from separate, well-formed bibliographic sections (Cioffi and Peroni, 2022). In contrast, it is still an unsolved problem to extract dispersed citation data, by which I mean information on references that is not centrally organized in a bibliography, but which might occur anywhere in the text of a scholarly work. For the purpose of the

2 In particular, ORCID is a self-maintained service and therefore cannot be used, for example, to retroactively identify deceased authors.

project, this mainly involves bibliographic information which is embedded in footnotes, often incomplete, mixed with commentary, and using heavily abbreviated terms. Fig 3 and 4 show examples of footnotes which demonstrate how different citation practices in the Humanities and Law are as compared to the highly structured reference section that can be found, for example, at the end of this chapter.

Fig. 2: Citation data on ZfRsoz in bibliometric data sources, as compared to the Journal of Law and Society

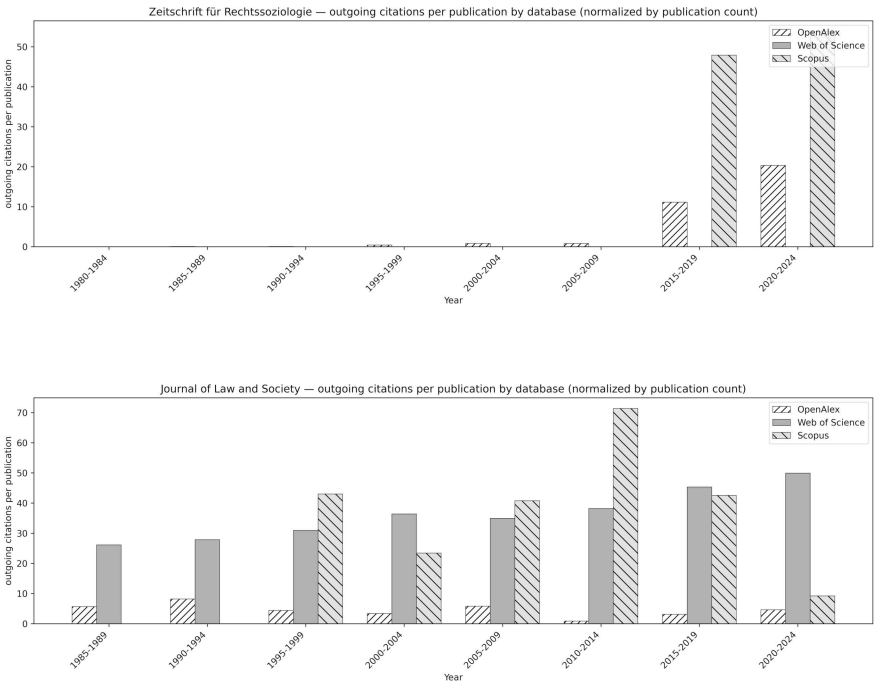


Fig. 3: Footnotes from a German historical article.

1 Ludwig Wittgenstein, Tractatus logico-philosophicus, London 1922, 1.1.

2 Vgl. die Artikel »Fact« im Oxford English Dictionary und »Thatsache« in Grimms Wörterbuch. Das Wort »T[h]atsache« scheint erstmals ins Deutsche eingeführt worden zu sein durch J. J. Spaldings Übersetzung von Bischof Joseph Butlers Schrift The Analogy of Religion, Natural and Revealed, to the Constitution and Course of Nature, 1736: W. Halbfass, Tatsache I, in: Historisches Wörterbuch der Philosophie, hg. von Joachim Ritter und Karlfried Gründer, 10 Bde., Darmstadt 1971 ff., Bd. 10, St–T, Sp. 909–914.

3 Halbfass, Tatsache (Fn. 2), Sp. 911.

4 Die Beziehungen zwischen den juristischen und den naturphilosophischen Schriften von Francis Bacon werden in der wissenschaftlichen Literatur immer stärker diskutiert. Zwei jüngere Studien, auf die ich mich stark gestützt habe, sind Julian Martin, Francis Bacon, the State and the Reform of Natural Philosophy, Cambridge 1992, und Barbara J. Shapiro, A Culture of Fact: England, 1550–1720, Ithaca, London 2000.

DOI: 10.12946/rg01/036-055

Fig 4: Footnotes from a German law journal article.

27 Vgl. den Beschluss der Kultusministerkonferenz: „Ausweitung der Nutzungsmöglichkeiten der Hochschulbibliotheken, ihre notwendige Ausstattung und verbesserte Verwendung ihrer Ressourcen" vom 27. Januar 1995, abrufbar unter: http://www.kmk.org/fileadmin/veroeffentlichungen_beschluesse/1995/1995_01_27-Nutzungsmoeglichkeiten-Hochschulbibliotheken.pdf, zuletzt abgerufen am 6.1.2020, S. 4: „Neue elektronische Formen der Fernleihe sind zunächst auf regionaler und nationaler Ebene zu entwickeln, dabei sind die Dienstleistungen zu beschleunigen und zu rationalisieren.".

28 Zu den verfassungsrechtlichen Hintergründen grundrechtsbezogener Optimierungsgebote, die hier nicht vertieft werden können, vgl.: *Böckenförde*, Der Staat 29 (1990), S. 21 f. et passim.

29 Dieser Sachverhalt wurde jüngst bestätigt in: EuGH, Urt. v. 19.12.2019 – C-263/18, ECLI:EU:C:2019:1111 = BeckRS 2019, 32135 – *Tom Kabinet*.

DOI: 10.5771/2699-1284-2020-1-16

Footnotes may contain several citations, often in one line. Also, many disciplines, such as law, heavily rely on abbreviated journal names, court cases, and shorthands. The challenge for extraction algorithms therefore is to find information embedded in a lot of noise. The existing tools, based on traditional machine-learning algorithms are not optimized for this kind of data and perform quite badly when faced with footnoted literature (Rodrigues Alves et al., 2018; Boulanger and Iurshina, 2022).

3. From structured prediction to large language models

This was the situation when I started work on the Legal Theory Knowledge Graph project in mid-2021.[3] The first goal was the implementation of a pipeline that would be able to process the pilot data: the PDFs of the "Zeitschrift für Rechtssoziologie" and "Journal of Law and Society". As no citation data exists for the "Zeitschrift", it quickly became apparent that the extraction of citations would be the main problem to solve, before any other part of the knowledge graph could be worked on.

At that point in time, LLMs were known only to a specialized audience and were not generally available. I experimented with several tools available at the time, which were all based on traditional machine learning techniques. GROBID (Testori, 2020) is a popular and well-maintained tool for the extraction of bibliographic data from PDFs. It is extremely well architected, using layered models that iteratively segment the document in order to extract structured bibliographic data. Ad-hoc tests showed, however, that it was not able to parse my source material correctly. Grobid provides ways of retraining it with domain-specific data. Unfortunately, the hierarchical layering of models that distinguishes Grobid from other tools also means that it requires highly complex training data. GROBID uses the TEI-XML format and a layered set of models and corresponding training datasets. Producing this data would have meant an investment of more time and resources than I could afford at the time.

3 https://www.lhlt.mpg.de/2514927

After experimenting with ExCite (Hosseini et al., 2019; Boulanger and Iurshina, 2022), I switched to using the AnyStyle citation extraction engine[4], which, like ExCite, relies on predictive machine learning and had shown to be, on average, the best performing tool in a comparison (Cioffi and Peroni, 2022). Similar to Grobid, the AnyStyle's default model had been trained with publications with structured bibliographies. AnyStyle's training data format is much simpler than Grobid's: AnyStyle uses only two models, one for extracting whole lines containing reference strings from text, and one for segmenting the references into its constituent parts.[5] Using a dedicated web application[6], and with the help of students, I produced a dataset of over a hundred XML annotations of PDF documents containing footnotes. Of these, 69 document annotations were used for training the extraction model and 882 fully annotated references for the segmentation model.[7] Using AnyStyle's own evaluation algorithm, we saw the following performance improvements:[8]

Fig. 5: Results of the evaluation using AnyStyle's check command

	Number of sequences	Sequence errors	Sequence error rate	Number of tokens	Token errors	Token error rate
Finder Model						
Default	69			57895	7033	**12.15%**
Custom	69			57895	1292	**2.23%**
Parser Model						
Default	882	734	**83.22%**	19428	7984	**41.10%**
Custom	882	170	**19.27%**	19428	1269	**6.53%**

Even with the quite limited number of annotations, we see a substantial decrease from a 12.15% error rate for the default extraction model that has been trained with non-footnoted literature to 2.23% for our custom dataset. For the segmentation model, the

4 https://github.com/inukshuk/anystyle.

5 The simplicity of AnyStyle comes with a cost: in contrast to GROBID, AnyStyle cannot profit from implicit semantic information provided by the coordinates of the page content: the placement helps to identify titles, authors, abstracts, or the footnotes, and differentiate these elements from other information.

6 In order to be able to involve non-technical support staff, I developed a web application that allows to visually annotate documents and produce training data in the AnyStyle format: https://github.com/cboulanger/citext. The application also allows using the currently trained model to automatically annotate the text loaded into the editor. The predicted annotations can be corrected and saved, to then re-train the model. This allows an iterative process with the aim to have a well-performing model that can deal with the complexity of footnote citations.

7 For this, the standard AnyStyle reference annotation label set was extended. In particular, the default segmentation model had problems with the fact that footnotes very often do not just contain reference data, but also "signal" words and phrases such as "see", "cf.", etc. and the author's commentary ("See, for example, the excellent overview by…") or other phrases that used to introduce sources ("As argued by …")(On this, see Vogel, 2012). Both types of content were annotated as "signal". This significantly improved the accuracy of reference segmentation.

8 The limitation of this algorithm is that it is testing the model against its gold standard data instead of cross-validating it with a split dataset. However, since the dataset is fairly heterogeneous, the results are meaningful for measuring the model's performance. Also, the check counts exact predictions only.

improvement is even more pronounced: the error rate goes down from 41.1% to 6.53%. However, even with this massive improvement, the extraction still did not yield results that matched our expectations. The resulting data was good enough to show some general trends, for example, to do a bibliometric analysis of a socio-legal journal (Boulanger, 2023; Boulanger, Creutzfeldt and Hendry, 2024). For these kinds of analyses, the errors in the data were acceptable, as we could expect them to be distributed over the feature space and therefore not to distort the results. However, for the purpose of populating a knowledge graph, the quality of the data was not sufficient.[9]

I had initially intended to publish the dataset[10] after adding more annotations and doing some quality control. However, after the "LLM revolution" at the end of 2022, it quickly became apparent that investing further time into training data for AnyStyle did not make sense. Ad-hoc experiments showed that even very early models such as OpenAI's "text-davinci-003" were outperforming the ML models without any training.[11] Therefore, I restarted the reference extraction project from scratch, relying on the ability of LLMs to infer the semantics of a reference instead of the "simple" pattern recognition afforded by traditional ML.

4. LLamore: Large LAnguage MOdels for Reference Extraction

In collaboration with David Carreto Fidalgo (mpcdf) and Andreas Wagner (mpilhlt), a new attempt was made to set up a workflow to extract citation data from the domain of (socio-)legal theory scholarship. Given the propensity of LLM models to hallucinate, it was clear that we needed a robust testing and evaluation solution before we could apply LLM at scale to analyse our data.

A cornerstone of this effort is the development of a high-quality "Gold Standard" dataset specifically tailored to Law and Humanities scholarship. In the previous attempts, the annotations were done on the source material that was of interest for the research question. The advantage was that the annotation efforts produced data that could directly be used for domain-specific analyses. A severe limitation of these documents was that they are not Open Access and therefore the source data, i.e. the PDFs, could not be distributed along with the Open Source code. We realized that if we wanted to advance the technology, in particular in collaboration with others, it was paramount to use Open Access journals to create a dataset. This would allow us to publish complete datasets including the source PDFs.

Therefore, we abandoned our previous data, and started again with an initial set of articles from Open Access journals. This dataset is currently being annotated using the

9 In particular, adding more training data did not seem to remove certain classes of parsing errors, such as the constant misclassification of "See … " as an author name instead of a signal word.

10 As the PDFs I was working with were not Open Access, for legal reasons, the dataset would have consisted only of raw footnote strings and the segmentation of these footnotes into references and the reference elements.

11 See https://pad.gwdg.de/s/LWUjPOi1F.

TEI XML standard.[12] TEI (Text Encoding Initiative) was chosen for its well-established, comprehensive framework for text interchange, which supports detailed markup beyond simple reference management, encompassing citations, cross-references, and contextual elements. While TEI offers robustness, it does not entirely resolve all conceptual or technical complexities inherent in representing these diverse citation practices.[13]

To leverage the capabilities of LLMs and evaluate their performance, we are using a lightweight Python package named "Llamore" (https://github.com/mpilhlt/llamore). Llamore facilitates two key functions: on the one hand, it can extract citation data from raw text or PDF inputs using various LLM APIs, which can be remote or locally hosted models, outputting structured data suitable for further processing. On the other hand, it provides methods for evaluating the accuracy of this extraction. Evaluation involves comparing the extracted references against the TEI-annotated Gold Standard, computing metrics such as F1-score by aligning the predicted elements with the gold elements.

Initial evaluation results comparing Llamore (using Gemini 2.0 Flash) with the established tool Grobid demonstrate the efficacy of the LLM-based approach for this specific domain, at least when using Grobid's default model. While Grobid and Llamore exhibit comparable performance on a standard dataset (F1 scores around 0.61-0.62)[14], Llamore significantly outperforms Grobid on the project's specialized dataset of footnoted SSH literature, achieving an F1 score of 0.45 compared to Grobid's 0.14. This represents approximately a threefold improvement in performance for handling complex footnotes.

To be sure, Grobid's model has not been trained with domain-specific material. In a cooperation with the current maintainer of Grobid, Luca Foppiano, we are annotating the same dataset for use by the Grobid models. Initial tests have shown that this has improved its performance, however, the production of a complete and big enough training dataset will take some time.[15] So the jury is still out as to whether LLMs or specialized tools will provide better performance, in particular since Grobid is much faster and consumes much less computing power. On the other hand, as we do not have a sufficiently large dataset yet, we have not been able to evaluate the performance of fine-tuned small LLMs.

5. Future work and outlook

These findings suggest that LLMs offer a viable path for extracting citation data from law and Humanities texts where traditional tools struggle. While Grobid remains a faster

12 We are using an online editor especially developed for this purpose. See https://github.com/mpilhlt/pdf-tei-editor.

13 See https://github.com/mpilhlt/bibliographic-tei/ for the documentation of cases where the TEI annotation guidelines do not provide clear rules for the markup of bibliographic information.

14 We used the "plos_1000" dataset provided by Grobid, published at https://zenodo.org/records/7708580.

15 See https://grobid.readthedocs.io/en/latest/Training-the-models-of-Grobid and https://grobid.readthedocs.io/en/latest/Grobid-specialized-processes/. In a cooperation with the University of Stuttgart, we are working on a set of high-quality annotations for training and evaluation.

and less resource-intensive option for the literature it was trained on, LLamore's performance on complex footnotes highlights the potential of modern language models. Future work will focus on expanding the Gold Standard dataset, refining evaluation metrics, and exploring the effectiveness of smaller, local, open-source models for this task. At the same time, the potential of solutions that do not rely on generative ai should be actively explored.[16]

Significant progress can only be made via open and collaborative efforts. First of all, the lack of training material for the citation practices in Law and the Humanities is one of the main reasons why current models perform badly. Since the production of high-quality training and evaluation data is notoriously difficult and time-consuming, we need a community effort to produce such material in amounts that will allow us to fine-tune and evaluate specialized models. In addition, the potential of synthetic datasets for finetuning needs to be explored.

Another stumbling block for collaboration concerns the lack of established standards for data interchange. We have decided to use TEI, which provides a syntax for expressing bibliographic information both as structured data and as annotated text, unlike other serialization formats such as BibTeX, MODS or CSL-JSON.[17] TEI provides a forward-compatible data format which allows to encode additional semantics such as the citation context or citation intent in future work.

In the context of knowledge graphs, a push towards more standardization, or at least for better interoperability, might come from the research on ontologies. Ontologies have been developed for bibliographic and citation data (Peroni and Shotton, 2018) or for data on scholars (Pertsas and Constantopoulos, 2017; Nguyen et al., 2020).[18] These ontologies might be useful for the eventual publication and long-term storage of the graph data produced in the project presented here.

In any case, much remains to be done before we have a sufficiently large, open knowledge graph that allows us to monitor and explore the evolution of a research field using data on knowledge production in the field of (socio-)legal theory. [19]

Acknowledgements

The data for fig. 2 were provided by the German Comptence Network for Bibliometrics funded by the Federal Ministry of Education and Research (Grant: 16WIK2101A)

16 See Wagner and Hermes (2026).

17 The corresponding TEI elements are <biblStruct> and <bibl>, respectively.

18 See https://opencitations.net/publications (last accessed 12.11.2025).

19 This chapter was written with support from large language models (LLMs). All model-generated text was reviewed and, where necessary, rewritten by the authors, who remain fully responsible for the final version. For details on the use of LLMs in this volume, see the statement in the volume's introduction.

References

Berrebbi D, Huynh N and Balalau O (2022) GraphCite: Citation Intent Classification in Scientific Publications via Graph Embeddings. In: 2nd International Workshop on Scientific Knowledge: Representation, Discovery, and Assessment. Available at: https://inria.hal.science/hal-03648498 (accessed 14 July 2025).

Boulanger C (2023) Citation graph of the Journal of Law and Society (1974–2022). DOI: 10.5281/ZENODO.8389925.

Boulanger C, Creutzfeldt N and Hendry J (2024) The Journal of Law and Society in context: a bibliometric analysis. Journal of Law and Society 51(1): 1–25. DOI: 10.1111/jols.12465.

Boulanger C, Fejzo, D and Rimmert, C (2025) Law Doesn't Count? Measuring Bibliometric Coverage of German Law Journals. Max Planck Institute for Legal History and Legal Theory Research Paper Series No. 2025–10. Available at http://dx.doi.org/10.2139/ssrn.5350481.

Boulanger C and Iurshina A (2022) Extracting Bibliographic References from Footnotes with Excite-Docker. In: Backes T, Iurshina A and Mayr P (eds) Proceedings of the Workshop on Understanding Literature References in Academic Full Text. Cologne: CEUR Workshop Proceedings, pp.26–33. Available at: http://ceur-ws.org/Vol-3220/#paper3 (accessed 14 July 2025).

Cioffi A and Peroni S (2022) Structured references from PDF articles: assessing the tools for bibliographic reference extraction and parsing. ArXiv. DOI: 10.48550/arXiv.2205.14677.

Cohan A et al. (2019) Structural Scaffolds for Citation Intent Classification in Scientific Publications. arXiv. DOI: 10.48550/arXiv.1904.01608.

Cohen DJ et al. (2008) Interchange: The Promise of Digital History. The Journal of American History 95(2): 452–491. DOI: 10.2307/25095630.

Gibson A and Ermus C (2019) The History of Science and the Science of History: Computational Methods, Algorithms, and the Future of the Field. Isis 110(3): 555–566. DOI: 10.1086/705543.

Gläser J and Oltersdorf J (2019) Persistent Problems for a Bibliometrics of Social Sciences and Humanities and How to Overcome Them. In: Catalano G et al. (eds) Proceedings of the 17th International Conference on Scientometrics & Informetrics. Rome: Edizioni Efesto, pp.1056–1567. Available at: https://www.issi-society.org/publications/issi-conference-proceedings/proceedings-of-issi-2019 (Accessed: accessed 14 July 2025).

Gottschalk S and Demidova E (2018) EventKG: A Multilingual Event-Centric Temporal Knowledge Graph. In: Gangemi A et al. (eds) The Semantic Web. Cham: Springer International Publishing, pp.272–287. DOI: 10.1007/978-3-319-93417-4_18.

Hammarfelt B (2016) Beyond Coverage: Toward a Bibliometrics for the Humanities. In: Ochsner M, Hug SE and Daniel H-D (eds) Research Assessment in the Humanities: Towards Criteria and Procedures. Cham: Springer International Publishing, pp.115–131. DOI: 10.1007/978-3-319-29016-4_10.

Haslhofer B, Isaac A and Simon R (2018) Knowledge Graphs in the Libraries and Digital Humanities Domain. In: Sakr S and Zomaya A (eds) Encyclopedia of Big Data Tech-

nologies. Cham: Springer International Publishing, pp.1–8. DOI: 10.1007/978-3-319-63962-8_291-1.

Hosseini A et al. (2019) EXCITE – A Toolchain to Extract, Match and Publish Open Literature References. In: 2019 ACM/IEEE Joint Conference on Digital Libraries (JCDL). Champaign: IEEE, pp.432–433. DOI: 10.1109/JCDL.2019.00105.

Jänicke S et al. (2015) On Close and Distant Reading in Digital Humanities: A Survey and Future Challenges. Eurographics Conference on Visualization (EuroVis) – STARs. DOI: 10.2312/EUROVISSTAR.20151113.

Jha R et al. (2017) NLP-driven citation analysis for scientometrics. Natural Language Engineering 23(1): 93–130. DOI: 10.1017/S1351324915000443.

Kuznetsov A (2019) Changed but Undescribed? What STS Could Say on the Research Practices of Social Sciences. EASST Review 38(1): 4–5. Available at: https://easst.net/easst-review/381/changed-but-undescribed-what-sts-could-say-on-the-research-practices-of-social-sciences/ (Accessed: 14 July 2025).

Liesegang L and Gläser J (2026) Supporting citation context analysis with large language models raises questions that should have been asked 40 years ago. In: Simons A, Wüthrich A, Zichert M, et al. (eds) *Understanding Science with Large Language Models? Potentials for the History, Philosophy, and Sociology of Science*. Bielefeld: transcript, part-6.

Lima JF, Amaral CMG and Molinaro LFR (2010) Ontology: An Analysis of the Literature. In: Varajão JEQ et al. (eds) ENTERprise Information Systems. Berlin: Springer, pp.426–435. DOI: 10.1007/978-3-642-16419-4_44.

Livermore MA and Rockmore DN (eds) (2019) Law as Data: Computation, Text, and the Future of Legal Analysis. Santa Fe: SFI Press.

Luhmann J and Burghardt M (2022) Digital humanities—A discipline in its own right? An analysis of the role and position of digital humanities in the academic landscape. Journal of the Association for Information Science and Technology 73(2): 148–171. DOI: 10.1002/asi.24533.

Machura S (2020) Milestones and Directions: Socio-Legal Studies in Germany and the United Kingdom. German Law Journal 21(7): 1318–1331. DOI: 10.1017/glj.2020.81.

Nguyen VB et al. (2020) Ontologies Supporting Research-related Information Foraging Using Knowledge Graphs: Literature Survey and Holistic Model Mapping. In: Keet CM and Dumontier M (eds) Knowledge Engineering and Knowledge Management, Springer International Publishing, 2020. DOI: 10.1007/978-3-030-61244-3_6.

Peroni S and Shotton D (2018) The SPAR Ontologies. In: Vrandečić D et al. (eds) The Semantic Web – ISWC 2018. Cham: Springer International Publishing, pp.119–136. DOI: 10.1007/978-3-030-00668-6_8.

Pertsas V and Constantopoulos P (2017) Scholarly Ontology: modelling scholarly practices. International Journal on Digital Libraries 18(3): 173–190. DOI: 10.1007/s00799-016-0169-3.

Rodrigues Alves D, Colavizza G and Kaplan F (2018) Deep Reference Mining From Scholarly Literature in the Arts and Humanities. Frontiers in Research Metrics and Analytics 3: 21. DOI: 10.3389/frma.2018.00021.

Simons A (2026) Scaling In, Not Up? Testing Thick Citation Context Analysis with GPT-5 and Fragile Prompts. In: Simons A, Wüthrich A, Zichert M, et al. (eds) *Understanding*

Science with Large Language Models? Potentials for the History, Philosophy, and Sociology of Science. Bielefeld: transcript, part-6.

Simons A, Arnaout H and Gurevych I (2026) Reconstructive citation context analysis using large language models. A roadmap. In: Simons A, Wüthrich A, Zichert M, et al. (eds) *Understanding Science with Large Language Models? Potentials for the History, Philosophy, and Sociology of Science*. Bielefeld: transcript, part-6.

Small H (2011) Interpreting maps of science using citation context sentiments: a preliminary investigation. Scientometrics 87(2): 373–388. DOI: 10.1007/s11192-011-0349-2.

Stichweh R (1992) 'The Sociology of Scientific Disciplines: On the Genesis and Stability of the Disciplinary Structure of Modern Science. Science in Context 5(1): 3–15. DOI: 10.1017/S0269889700001071.

Testori M (2020) GROBID: when data extraction becomes a suite. In: OpenMethods. Available at: https://openmethods.dariah.eu/2020/09/09/grobid-when-data-extraction-becomes-a-suite/ (accessed 14 July 2025).

Underwood T (2017) A Genealogy of Distant Reading. Digital Humanities Quarterly 11(2). Available at: http://digitalhumanities.org:8081/dhq/vol/11/2/000317/000317.html (accessed 14 July 2025).

Vera-Baceta MA, Thelwall M and Kousha K (2019) Web of Science and Scopus language coverage. Scientometrics 121(3): 1803–1813. DOI: 10.1007/s11192-019-03264-z.

Vogel R (2012) Verbs for referring to sources in humanities and social sciences: Grammatical and lexical analysis of their distribution. Discourse and Interaction 5(1): 63. DOI:10.5817/DI2012-1-63

Wagner A and Hermes J (2026) Encoded humanities, or: not everything has to be generative. A dialogue on AI tasks and roles. In: Simons A, Wüthrich A, Zichert M, et al. (eds) *Understanding Science with Large Language Models? Potentials for the History, Philosophy, and Sociology of Science*. Bielefeld: transcript, part-1.

Zhang F et al. (2021) RDF for temporal data management – a survey. Earth Science Informatics 14(2): 563–599. DOI: 10.1007/s12145-021-00574-w.