Published in: DeGaetano, Stefania/Kazantseva, Anna/Reiter, Nils/Szpakowicz, Stan (Eds.): Proceedings of the
The 4th Joint SIGHUM Workshop on Computational Linguistics for Cultural Heritage, Social Sciences, Humanities and Literature.
December 12, 2020, Barcelona, Spain (Online). - Stroudsburg: Association for Computational Linguistics, 2020. Pp. 112-120.
URL: *https://www.aclweb.org/anthology/2020.latechclfl-1.13/*

Twenty-two Historical Encyclopedias Encoded in TEI: a New Resource for the Digital Humanities

Thora Hagen
University of Würzburg
thora.hagen@uni-wuerzburg.de

Erik Ketzan
University of Cologne
eketzan@uni-koeln.de

Fotis Jannidis
University of Würzburg
fotis.jannidis@uni-wuerzburg.de

Andreas Witt
IDS Mannheim & University of Cologne
andreas.witt@uni-koeln.de

Abstract

This paper accompanies the corpus publication of EncycNet, a novel XML/TEI annotated corpus of 22 historical German encyclopedias from the early 18th to early 20th century. We describe the creation and annotation of the corpus, including the rationale for its development, suggested methodology for TEI annotation, possible use cases and future work. While many well-developed annotation standards for lexical resources exist, none can adequately model the encyclopedias at hand, and we therefore suggest how the TEI Lex-0 standard may be modified with additional guidelines for the annotation of historical encyclopedias. As the digitization and annotation of historical encyclopedias are settling on TEI as the de facto standard, our methodology may inform similar projects.

1 Introduction

EncycNet is a TEI-annotated corpus of 22 historical German encyclopedias from the early 18th to early 20th century, over 49,300,000 word tokens, with the goal of providing a resource for NLP and to add to the growing amount of historical, lexicographical German texts in consistently annotated XML. Initial versions of the texts were provided to us by Zeno.org in a proprietary XML schema, and we then developed a TEI annotation schema that fits the diverse structures of these encyclopedias and at the same time compromises with existing encoding standards with regard to annotating lexicographic works. We then applied the annotation schema to our corpus using XSLT (Extensible Stylesheet Language Transformations, a language for transforming XML documents into other XML documents).

The TEI methodology described in detail below is intended to connect our corpus to existing annotated, lexicographic corpora (while still accurately representing the original entry structures), as well as the semantic web, and also provide similar projects working with historic encyclopedias, in a variety of languages, with a reference point.

In this paper, we first comment on the value of encyclopedia texts for humanities research, summarize related work, and provide an overview of the individual encyclopedias included in this corpus. We then present our choices for TEI annotation and other transformation objectives that intend to homogenize the data and improve shareability and incorporation into related projects. Specific encoding choices and corresponding examples from the corpus will be given when necessary. The encyclopedia corpus is openly accessible under a CC-BY license on Zenodo[1][2] while the XSLT files are available on GitHub.[3]

Following corpus publication, we will begin the creation of EncycNet as a knowledge graph, which we hope will aid research in diachronic linguistic change in the German language, enriching existing

[1]Original corpus: http://dx.doi.org/10.5281/zenodo.4159491

[2]Transformed corpus, including the ODD: http://dx.doi.org/10.5281/zenodo.4039569

[3]https://github.com/ThoraHagen/Encyc-Transformation

Proceedings of LaTeCH-CLfL 2020, pages 112–120
Barcelona, Spain (Online), December 12, 2020.
Publikationsserver des Leibniz-Instituts für Deutsche Sprache
URN: http://nbn-resolving.de/urn:nbn:de:bsz:mh39-102512

language models with explicit semantic knowledge, and humanistic examination of changes in societal attitudes over time, with centuries of aligned encyclopedia texts as evidence.

2 Related Work

2.1 Encyclopedias

Encyclopedias occupy an important role in cultural history as texts which profess to aim, naturally in manifold ways, to encapsulate the entirety of human knowledge or a specific field within a systematic form. Regardless of how one reads or responds to them, statements such as Thorndike's (1924) assertion that encyclopedias are "the most important monuments of the history of science and civilization" are profuse in the history of cultural commentary, with just one more recent example being Rosenberg's (1999) assertion that "the importance of [Diderot's *Encyclopedia*] to the Enlightenment is difficult to overstate". Encyclopedias offer scholars a wealth of text for scholars of fields too numerous to name, not only due to the many fields of knowledge covered, but also to the variety of formal and conceptual structure, including, in the German-speaking world, the *Konversationslexikon* or "conversation dictionary", intended to contain "general knowledge" for various audiences. But, as Belgum (2010) notes, "The problem with encyclopedias for the scholar [...] is the vast amount of information and commensurately large number of topics they contain", suggesting that digital humanities methods may be useful in assisting linguists, historians, and other cultural scholars in contending with the sheer length of the texts. And indeed, DH investigations of the content of encyclopedias have emerged (e.g. Seifert (2007); Hagen (2007)).

While most notable historical encyclopedias are available on the Web in some form, often as scans or plain text, many suffer from restricted access, issues of long term archiving, and sometimes unclear licensing. Some annotated encyclopedias are restricted behind paywalls, e.g. Diderot's canonical encyclopedia digitized by ARTFL project,[4] which is queryable through a web interface but requires paid subscription for full text files. Other projects have created impressive and user-friendly online interfaces for encyclopedia texts but did not share corpora openly online, e.g. the University of Trier's Krünitz Online[5] (which has deposited XML/SGML-annotated data with the University Library of Trier) and University of Heidelberg's "Hidden Grammars of Transculturality – Migrations of Encyclopaedic Knowledge and Power"[6] and "Encyclopedia Database",[7] which followed the TEI markup guidelines described by team member (Petersen, 2010). The Encyclopedia Database website no longer seems to be functional, underscoring the risk that research outputs may be rendered unusable when long-term archiving is not successful. While the Encyclopedia Database website states that data is "open-source [... and we] invite all non-commercial usage of our data, and encourage cooperation partners to exchange their materials with us", the project website also suggests that some resources still under copyright or otherwise restricted were used in the project,[8] leading us to question what rights or licenses may be attached to such data, even if it may still be obtained. EncycNet aims to improve upon this state of the art through: open access to the corpus, use of the widely-used and unrestrictive CC-BY license, long term archiving at Zenodo, and the use of a TEI standard which other projects may follow.

2.2 Choice of Annotation Standard

While TEI has become the de facto standard for annotation of lexical and related resources, there is no widespread de facto methodology for TEI encoding of encyclopedia texts, as the previous section illustrates. TEI P5 provides the module 'Dictionaries' for annotating any terminological or lexicographical texts, but its rules are designed to fit such a wide variety of these types of texts that they can be considered

[4] ARTFL Encyclopédie, https://encyclopedie.uchicago.edu/

[5] Krünitz Online, http://www.kruenitz1.uni-trier.de/

[6] Hidden Grammars of Transculturality – Migrations of Encyclopaedic Knowledge and Power, https://www.asia-europe.uni-heidelberg.de/en/research/d-historicities-heritage/d11.html

[7] Encyclopedia Database, http://kjc-sv036.kjc.uni-heidelberg.de:8080/exist/apps/matumi/home.html

[8] Encyclopedia Database, "Select encyclopedias and articles for digitization", http://kjc-sv036.kjc.uni-heidelberg.de:8080/exist/apps/matumi/home-history.html

too general (see section on TEI Lex-0 below). In this section, we briefly introduce the available choices for annotating encyclopedias and explain why TEI Lex-0 is the appropriate choice for EncycNet.

TEI has become a widespread standard for text annotation, but terminology poses a special case as already mentioned. Until P4, the TEI included a module on terminology, which was removed with P5 in 2007 as the module was deemed obsolete (due to the publication of related ISO standards). This sparked a discussion on how to best handle terminology along with the proposal of many new annotation guidelines (Van Campenhoudt, 2017). Besides simply utilizing the core elements and the module 'Dictionaries' of TEI P5 (Budin et al., 2012), there now exist a few standards for annotating lexicographic / terminographic data, most notably TBX, TEI-TBX and TEI Lex-0. Drawing specifically on previous work in annotated encyclopedias, the most extensive TEI schema was suggested by Petersen's "A Minimal Set of Tags for Marking Up Encyclopaedias" (2010), which is now mostly covered by newer standards. However, Petersen also discusses a few points, such as footnotes in entries or the importance of paragraphs, that have still not been addressed in commentary on TEI for lexicographic works (see also section 4.1).

One ISO standard leading to the removal of the terminology module is TBX (TermBase eXchange),[9] which has become a popular choice for encoding terminology. TBX is entirely separate from TEI and its dialects, even if the proposed document structure is strongly inspired by the TEI, and was adopted by ISO in 2008. TBX is flexible, can express almost any kind of terminological data, and (as with TEI) encompasses a few dialects to better fit certain needs, such as TBX-Basic or TBX-Min. TBX is generally intended to be used for onomasiological approaches, i.e. dictionaries that group all synonymous terms together based on meaning (Romary and Witt, 2014).

To compensate for the fact that TBX is not compatible with TEI, a new project called TEI-TBX (Romary, 2014) emerged to combine the two. The main idea behind the linking of TBX and TEI is to integrate this onomasiological approach back into TEI by only taking the TBX-Basics representation of an entry into account to compromise between the two standards.

A recent encoding schema supported by DARIAH is TEI Lex-0 (Romary and Tasovac, 2018), which is based on the idea of further specifying the 'Dictionary' module of TEI, especially to fit lexicographic works better. The guidelines are much more straight-forward, as more restrictions on existing tags were introduced and certain tags got removed entirely. Because these guidelines are only a customization (not a replacement) of the TEI schema, any TEI Lex-0 valid document is therefore also TEI valid. This schema, as already implied above, enforces the semasiological model (synonymous terms are grouped in alphabetical order (Romary and Witt, 2014)) as an option besides the onomasiological model, compared to TBX. As the project is community-oriented, the standard is continually updated to fit the needs of its users. Because encyclopedias follow the semasiological approach and these guidelines provide a more explicit frame than base TEI does, we chose TEI Lex-0 as our foundation for annotating EncycNet's texts.

3 Data

The original texts of the twenty-two German encyclopedias in our corpus were initially made available to us by Zeno as part of a larger, literary corpus for an earlier data transformation / preparation project. The focus of this previous project was a wide variety of literary text forms, meaning that the transformation process was designed to fit a wider variety of ambiguous XML markup (i.e. not specifically tailored for the encyclopedias which made up only part of their larger corpus). In EncycNet, we are now able to attend to the specific structures and markup of encyclopedias texts while re-evaluating previous transformation decisions, most importantly choosing a TEI P5 `<teiCorpus>` approach. For a more detailed view of the content and the size of the corpus, see Table 1.

The EncycNet corpus has also been annotated with semantic web data, namely the automatic alignment of encyclopedia headwords with DBpedia (Auer et al., 2007), which contains millions of structured and classified entities from Wikipedia, and GermaNet, a lexical-semantic net for the German language (Hamp and Feldweg, 1997; Henrich and Hinrichs, 2010), together with its aligned Wiktionary definitions (Henrich et al., 2011). DBpedia Spotlight (Daiber et al., 2013) provides a disambiguation algorithm,

[9] TBX, `https://www.tbxinfo.net/`

	Description	Number of Entries	Token Count
Brockhaus 1809	General lexicon targeted at the general population.	6,960	1,186,000
Brockhaus 1837	General lexicon briefly covering everyday subjects in a strictly non-scientific way, while focusing on illustrations.	7,049	2,604,000
Brockhaus 1911	General, pocketbook edition lexicon.	82,780	2,434,000
DamenConvLex. 1834	General encyclopedia explicitly targeted at middle-class women interested in education.	7,099	1,461,000
Eisler 1904	Dictionary covering philosophical terms.	3,646	845,000
Eisler 1912	Lexicon covering influential philosophers.	2,839	396,000
Goetzinger 1885	Specialist encyclopedia on the cultural history of the German people based on the existing scientific research at the time.	948	519,000
Hederich 1770	Lexicon of mythology, addressed to scholars and artists.	6,430	832,000
Heiligenlex. 1858	Biographies of saints of the catholic church.	33,481	3,095,000
Herder 1854	General lexicon with short explanations of various topics.	39,755	2,256,000
Kirchner-Michaelis 1907	Dictionary briefly covering basic philosophical terms, including their historical development.	1,817	235,000
Lemery 1721	Lexicon covering minerals, animals, and herbs. Mostly targeted at doctors, pharmacists, etc., but also at the general population.	1,769	495,000
Lueger 1904	Specialist lexicon covering technical terms.	23,465	5,246,000
Mauthner 1923	Dictionary of philosophical terms.	214	537,000
Meyers 1905	Comprehensive general lexicon targeted at the general population.	156,264	17,437,000
Pagel 1901	Biographical lexicon of important doctors of the 19th century.	2,909	528,000
Pataky 1898	Lexicon of female, German authors, covering their biographies, works and pseudonyms.	6,907	503,000
Roell 1912	Encyclopedia of the railway industry.	3,067	2,662,000
Schmidt 1902	Lexicon of the history of German booksellers and book printers.	567	380,000
Sulzer 1771	Encyclopedia of terms within the field of aesthetics.	854	816,000
Vollmer 1874	Dictionary of mythological terms and their development in different cultures.	7,078	479,000
Wander 1867	Dictionary of German proverbs.	25,762	5,242,000

Table 1: Short description as well as the number of entries and overall token count of each encyclopedia in the corpus.

while the Lesk algorithm (Lesk, 1986) was used for GermaNet.[10] In both cases, the encyclopedia glosses were used as context input for disambiguating the encyclopedia headword in semantic web classification. Table 2 shows the percentage of headwords which were able to be automatically aligned with DBpedia and GermaNet in this way, and conversely, provide estimations of the amount of knowledge in our corpus that is currently *not* part of these semantic web resources. Thus, our corpus could be used to further add entries to DBPedia and GermaNet in the future.

The original encyclopedias encode one encyclopedia entry in one `<article>` element, while the child element `<lem>` contains the headword and the child element `<text>` contains the gloss itself. The usage of these elements is, along with typography elements such as `<i>` and `<u>`, fairly consistent and unambiguous. However, the internal structure of a gloss can be vastly differently annotated. Most notably, `<p>` can indicate a paragraph, a line from a verse, or a list item. `<p>` elements can have an *@class* attribute assigned, but these classes do not align with their structural function, sometimes not even within the same encyclopedia. The same problem also applies to references. Apart from the issue that references are not explicitly linked, the elements `<link>`, `<plink>`, and other typographical elements can indicate a footnote, gloss reference or footnote "anchor"—and these elements are also used interchangeably. In section 4, these two main issues (basic structure and references) in the transformation

[10]We used DBpedia Spotlight because it is specifically tailored to DBpedia. As no such tailored tool exists for GermaNet, we opted for Lesk as the algorithm and its variants still belong to the standards of German Word Sense Disambiguation (Henrich and Hinrichs, 2012). Other tools for entity linking such as spaCys Entity Linker or DeepType (Raiman and Raiman, 2018) could have been used besides the approaches here; however at this early stage of the project our current claims do not benefit from marginal performance improvements.

	GermaNet (%)	DBpedia (%)	Internal Alignment (%)
Brockhaus 1809	32.80	51.15	71.49
Brockhaus 1837	58.04	54.38	91.84
Brockhaus 1911	24.80	37.18	73.68
DamenConvLex. 1834	47.68	43.61	70.70
Eisler 1904	46.05	25.40	57.87
Eisler 1912	10.21	5.92	21.86
Goetzinger 1885	56.54	48.31	76.26
Hederich 1770	4.45	9.38	26.26
Heiligenlex. 1858	2.34	0.35	1.10
Herder 1854	27.25	34.61	67.32
Kirchner-Michaelis 1907	58.56	29.66	75.23
Lemery 1721	7.91	18.65	41.09
Lueger 1904	28.77	24.57	46.51
Mauthner 1923	75.23	28.97	77.10
Meyers 1905	19.80	28.94	47.03
Pagel 1901	11.28	0.17	19.09
Pataky 1898	15.43	0.58	18.00
Roell 1912	24.91	27.88	35.78
Schmidt 1902	12.70	1.06	9.47
Sulzer 1771	69.67	24.71	72.01
Vollmer 1874	7.49	14.51	34.84
Wander 1867	43.75	16.06	33.27

Table 2: Percentage of encyclopedia headwords in EncycNet for which a GermaNet synset entry or DBpedia URI was automatically assigned. The last column additionally shows the amount of headwords appearing at least once in the remaining 21 encyclopedias.

will be discussed in more detail.

A markup conversion was therefore necessary to make working with these encyclopedia texts on a quantitative scale possible. For the XSL transformation process, we focused on preserving all data as a general rule. To ensure that no text instances were lost, we implemented a quantitative evaluation, i.e. compared the glosses of all original encyclopedias with their counterpart from their respective transformed encyclopedia with Python. We also carried out a qualitative evaluation by transforming the results into simple HTML and randomly manually comparing parts with their facsimile equivalent.

4 Modifying the TEI Lex-0 Guidelines and Transformation Objectives

4.1 Basic Structure

TEI Lex-0 is an actively maintained annotation schema for lexical resources and fits most needs of our corpus, including in TEI Lex-0's semasiological approach (term: definition), which relates to the canonical structure of encyclopedias, including our German ones.

The glosses of the encyclopedias do not exhibit a typically lexicographical structure such as *headword, grammatical features (and possibly many other kinds of features), short definition, usage example*. We nonetheless chose to adopt the predominant encoding style that exists for lexicographic data (`<entry>` structure) instead of opting for a more general approach (e.g. a `<teiCorpus>` structure), so that researchers working with lexicographic data might be able to integrate these encyclopedias into their workflow more easily.

An example of a minimal gloss in our encyclopedia corpus:

```
<?xml version="1.0" encoding="UTF-8"?>
<entry xml:id="d3e45097Brockhaus-1809" xml:lang="de">
   <form type="lemma">
      <term>Die Literatur</term>
   </form>
   <sense xml:id="d3e45097">
      <def>
         <term type="headword"><hi rend="bold">Die Literatur</hi>
         </term>, a. d. Lat. 1) im Allgemeinen die Gelehrsamkeit; 2) insbesondre
              pflegen Einige dieses Wort auf die <hi rend="italic">schoenen
            Wissenschaften</hi> einzuschraenken.
     </def>
   </sense>
</entry>
```

Besides typography-focused elements such as <hi>, there are also structure-focused elements that can be frequently found within the glosses, namely tables, verse, and lists. Within TEI Lex-0 valid documents, these structures are not allowed. We therefore slightly modified the guidelines, so that glosses may contain these elements, as well.

We also decided on another TEI modification decision concerning the content of <def> elements. In the original encyclopedia files, glosses are composed of multiple paragraphs, realized with <p> elements. The paragraphs cannot be directly transformed into <def> elements, because one paragraph usually does not align with one definition. As we did not want to lose the paragraph structure by compiling all adjacent paragraphs into one <def> element, we chose to allow <p> elements within <def> elements instead, which is not compliant with TEI P5. We argue, however, that only allowing phrase-like elements within a definition is too restrictive for encyclopedias in general, as it is possible for a definition to consist of multiple paragraphs. For our encyclopedia structure this means that <sense> elements are always composed of just one <def> element containing either two or more <p> elements or the definition directly.

Footnotes are an essential aspect of this corpus, but so far, while <note> is generally allowed, no explicit guidelines for handling footnotes in encyclopedias exist in a standardized form yet. Petersen (2010) suggests replacing all footnote references with the corresponding footnote text, however footnotes with no references still need to be addressed then and footnotes with many references pointing to them might unnecessarily inflate the text. We instead chose to include any footnotes at the end of one <def> element as a <note> element, so that in case of two or more definitions, their corresponding footnotes will be grouped with them. Single footnotes are again annotated with a <note> with an attribute *@type* as 'footnote'. In these encyclopedias, there exist cases where footnotes are not placed at the end of entries. In the *Heiligenlexikon*, for example, all footnotes are grouped at the end of one letter range. In Wander's *Sprichwörter-Lexikon*, some entries contain notes in the middle of definitions. In both cases, we moved all footnotes to the end of their respective entries if necessary, thus altering the original structure to preserve a uniform annotation. We thus suggest that: all footnotes, like entries, should carry a unique *@xml:id* identifier, should only be allowed at the end of the related definition, and should carry the aforementioned *@type* attribute to distinguish them from other notes.

4.2 Encoding References and Footnotes

Another challenge in annotating the encyclopedias was to create explicitly tagged links between references, for instance when an encyclopedia gloss contains the text *see also X* or *see external source Y*. We thus assigned reference targets with a unique identifier as a first step and five different types of references are present in the encyclopedia corpus: references to glosses, footnotes, figures, external sources, and appendices.

We first want to focus on gloss references. Previously, some references were already tagged with fairly unambiguous elements such as <link>; others are simply highlighted with elements such as <i>.

We tried to identify highlighted references next to ordinarily annotated references with distinct regular expressions as a second step, so that phrases such as *as seen in article* *rhythm* (translated from entry *Der Takt* in Brockhaus' *Conversations-Lexikon*) can be resolved. Following the identification procedure, every such reference is then annotated with <ref> and an attribute *@type* with the value 'entry'. Connecting all recognized entry references to their actual entry identifiers constitutes the last step. As an indication to what entry a reference should be linked to is firstly any attribute values from elements like <link> and secondly the text content of the element. Usually, either one or the other string matches exactly one entry headword. We also implemented additional rules like substring matching or switching of first and last names for persons to resolve as many references as possible.[11] However when both strings are either too ambiguous (e.g. when just the first name is given for a reference to a person), the transformation will result in no target for that entry. This can also happen when the text content irregularly contains superfluous information or is spelled differently then the targeted entry headword (and no attribute is available), such as *biography see Anna v. Gottberg née baroness v. Rottenberg* in Patakys *Deutsches Lexikon deutscher Frauen der Feder*. Sometimes entries of references could not be located even by manual search. To give an example from Brockhaus 1837: 5,644 entry references could be found from which 4,907 match a unique entry identifier. Previously, no entry references were explicitly tagged in the original XML document. References to appendices are marked as such via *@type*, but are identified and resolved in the same way, with the exception of taking 'appendix markers' (e.g. quotation marks in Meyers *Großes Konversations-Lexikon*) into account. External references in this corpus can only be identified with regular expressions and are thus far left without a target.

Two encyclopedias contain figure references: Rölls *Enzyklopädie des Eisenbahnwesens* and Lügers *Lexikon der gesamten Technik*. These references are realized with <anchor> elements, as the references point to different parts of the figure, and consequently require different anchors and identifiers for the same figure. We chose to collapse all identifiers pointing to the same figure onto the same figure identifier and remove the anchors, as the redundant pointing does not make sense from a digital perspective. The connection between the figure part and the reference remains intact on a human-readable basis via the references' text content.

Lastly, there exist footnote references. In the original files, footnotes were only unique within the context of one entry, but after the transformation, footnotes were assigned a distinctive identifier by concatenating the headword with the footnote number, as homonymous headwords are also numbered. For each footnote reference, the footnote with the identical number within the same entry is assigned as a target.

5 Conclusion and Future Work

With the publication of EncycNet's newly annotated encyclopedia corpus, we hope to provide a novel, digital, historical resource and advance the methodology of TEI for encyclopedia texts. We argue that encyclopedias such as these should be treated just like any other lexical resource and thus should be annotated as such, even when the gloss structure they exhibit does not follow typical standards like any other lexical resource. Our proposed methodology, a compromise between TEI and TEI Lex-0, can be summarized as follows: allow a paragraph-like structure in definitions, allow less common structures such as lists and tables in definitions, and create explicit rules for footnotes within entries. We strive to include these findings within a future, revised version of the TEI and TEI Lex-0.

While our next step in this project is the transformation of the EncycNet corpus into a knowledge graph structure, the corpus at this earlier stage may still be improved. Currently, many headword features such as synonyms, hyper- or hyponyms are marked with highlight-elements, for example. With TEI Lex-0, these features could theoretically be annotated as such. Automatically distinguishing tagged synonyms from other words marked with the same highlight tag is possible, but not realistic within the scope of an XSL transformation, which is why these features remain "unidentified" for now. A future version

[11] A peer reviewer stated that a confidence score for reference and footnote matches based on e.g. the preciseness of the rule (full string vs. substring) that applied for all individual matches could be of interest. As it is possible to include certitude values as an attribute from TEI's viewpoint, we would like to add this feature in a future version of the corpus.

of the corpus, where features have been explicitly tagged with the help of appropriate NLP methods and another programming language, is still conceivable. Another unresolved issue is the grouping of multiple senses in one <sense> element, which can sometimes be the case for homonyms, but also entries listing multiple persons from the same family for example. Ideally, these senses should be split into two <sense> tags within the same <entry>, as TEI Lex-0 proposes. Identifying and splitting such glosses is, similarly to identifying headword features within the definition, difficult to realize with XSLT alone but feasible with other tools.

We hope that EncycNet, in both its TEI-annotated corpus and planned RDF conforming knowledge graph form, may aid a wide variety of future research. EncycNet adds a large new resource for historical German, and diachronic language change as well as language models may use EncycNet as a source of additional historical and semantic knowledge. Given the unique role of encyclopedias in cultural history, EncycNet also offers a wealth of evidence for humanistic research. As shown in the alignment of headwords in Table 2 above, the encyclopedia texts have a broad range of overlap in subject matter, and much work could be done on, e.g., what topics different encyclopedias include and exclude, the length of entries for topics, etc., as evidence of societal attitudes of the times. As just one example, the entries in the 1834 "women's" encyclopedia could be compared to general knowledge encyclopedias of the same time period to better understand the editors' and, arguably, the contemporary society's social prejudices and biases against women. Finally, we expect that openly published, TEI-encoded encyclopedia projects will continue to emerge in a variety of languages; one recent example is the ongoing "Nineteenth-Century Knowledge Project" (Logan, 2018; Grabus et al., 2019) which digitizes historic editions of the *Encyclopedia Britannica*. We suggest that our TEI methodology could be one way that such projects may benefit from one another and enrich our understanding of encyclopedia texts.

References

Sören Auer, Christian Bizer, Georgi Kobilarov, Jens Lehmann, Richard Cyganiak, and Zachary Ives. 2007. DBpedia: A Nucleus for a Web of Open Data. In Karl Aberer, Key-Sun Choi, Natasha Noy, Dean Allemang, Kyung-Il Lee, Lyndon Nixon, Jennifer Golbeck, Peter Mika, Diana Maynard, Riichiro Mizoguchi, Guus Schreiber, and Philippe Cudré-Mauroux, editors, *The Semantic Web*, pages 722–735, Berlin, Heidelberg. Springer Berlin Heidelberg.

Kirsten Belgum. 2010. Documenting the Zeitgeist How the Brockhaus Recorded and Fashioned the World for Germans. *Publishing Culture and the 'Reading Nation'. German Book History in the Long Nineteenth Century*, pages 89–117.

Gerhard Budin, Stefan Majewski, and Karlheinz Mörth. 2012. Creating Lexical Resources in TEI P5. A Schema for Multi-purpose Digital Dictionaries. *Journal of the Text Encoding Initiative*, (3).

Joachim Daiber, Max Jakob, Chris Hokamp, and Pablo N. Mendes. 2013. Improving Efficiency and Accuracy in Multilingual Entity Extraction. In *Proceedings of the 9th International Conference on Semantic Systems (I-Semantics)*.

Sam Grabus, Jane Greenberg, Peter Logan, and Jane Boone. 2019. Representing Aboutness: Automatically Indexing 19th-Century Encyclopedia Britannica Entries. *NASKO*, 7(1):138–148.

Nadine Hagen. 2007. *A mind-map of a nation: the Australian encyclopaedia or why sharks are more important than tigers*. Shaker Verlag.

Birgit Hamp and Helmut Feldweg. 1997. Germanet-a lexical-semantic net for german. In *Automatic information extraction and building of lexical semantic resources for NLP applications*.

Verena Henrich and Erhard Hinrichs. 2010. GernEdiT-the GermaNet editing tool. In *Proceedings of the ACL 2010 System Demonstrations*, pages 19–24.

Verena Henrich and Erhard Hinrichs. 2012. A Comparative Evaluation of Word Sense Disambiguation Algorithms for German. In *LREC*, pages 576–583.

Verena Henrich, Erhard Hinrichs, and Tatiana Vodolazova. 2011. Aligning GermaNet senses with Wiktionary sense definitions. In *Language and Technology Conference*, pages 329–342. Springer.

Michael Lesk. 1986. Automatic sense disambiguation using machine readable dictionaries: how to tell a pine cone from an ice cream cone. In *Proceedings of the 5th annual international conference on Systems documentation*, pages 24–26.

Peter Melville Logan. 2018. Nineteenth-Century Knowledge Project. *JADH 2018*, page 226.

Jens Østergaard Petersen. 2010. A Minimal Set of Tags for Marking Up Encyclopaedias.

Jonathan Raiman and Olivier Raiman. 2018. Deeptype: multilingual entity linking by neural type system evolution. *arXiv preprint arXiv:1802.01021*.

Laurent Romary and Toma Tasovac. 2018. TEI Lex-0: A Target Format for TEI-Encoded Dictionaries and Lexical Resources. In *TEI Conference and Members' Meeting*, Tokyo, Japan, September.

Laurent Romary and Andreas Witt. 2014. Méthodes pour la représentation informatisée de données lexicales / Methoden der Speicherung lexikalischer Daten. *Lexicographica*, 30(1):152–186.

Laurent Romary. 2014. TBX goes TEI–Implementing a TBX basic extension for the Text Encoding Initiative guidelines. *arXiv preprint arXiv:1403.0052*.

Daniel Rosenberg. 1999. An Eighteenth-Century Time Machine: The" Encyclopedia" of Denis Diderot. *Historical Reflections/Réflexions Historiques*, pages 227–250.

Hans-Ulrich Seifert. 2007. Dewey meets Krünitz. Semi-Automized Classification in Historical Encyclopaedias. pages 95–104.

Lynn Thorndike. 1924. L'Encyclopedie and the History of Science. *Isis*, 6(3):361–386.

Marc Van Campenhoudt. 2017. Standardised Modelling and Interchange of Lexical Data in Specialised Language. *Revue française de linguistique appliquée*, 22(1):41–60.