Umanistica Digitale – ISSN: 2532-8816 – n. 23, 2026: Special Issue
H. Leithe-Jasper, R. Marchionni, N. Götze – *Mittellateinisches Wörterbuch and Thesaurus linguae Latinae – jointly addressing digital challenges*
DOI: https://doi.org/10.60923/issn.2532-8816/23769

Mittellateinisches Wörterbuch and Thesaurus linguae Latinae – jointly addressing digital challenges [1]

Helena Leithe-Jasper

Österreichische Akademie der Wissenschaften
h.leithejasper@mlw.badw.de

Roberta Marchionni

Bayerische Akademie der Wissenschaften Muenchen, Thesaurus linguae Latinae
marchionni@thesaurus.badw.de

Nora Götze

Bayerische Akademie der Wissenschaften
nora.goetze@badw.de

Abstract

Il *Thesaurus linguae Latinae* e il *Mittellateinisches Wörterbuch*, entrambi progetti con sede presso l'Accademia Bavarese delle Scienze di Monaco, si occupano del lessico latino e coprono insieme un arco temporale che va dalle prime attestazioni scritte fino alla fine del XIII secolo. Anche per questi progetti di lunga tradizione, le nuove tecniche di digitalizzazione indicano la via verso cambiamenti profondi, necessari e fruttuosi, che riguardano il workflow, la gestione dell'ingente materiale e l'accessibilità globale dei risultati. Dopo aver valutato diverse proposte, sono state adottate varie soluzioni per ciascuno di questi ambiti: alcune sono già operative, mentre altre sono attualmente in fase di implementazione. Quanto segue è il resoconto di tre collaboratrici.

Parole Chiave: lessicografia; banca dati; DSL (domain-specific language); biblioteca.

[1] Section 1 is authored by Helena Leithe-Jasper; section 2 is authored by Roberta Marchionni; section 3 is authored by Nora Götze.

The Thesaurus linguae Latinae *and the* Mittellateinisches Wörterbuch*, both based at the Bavarian Academy of Sciences and Humanities in Munich, deal with the Latin lexicon and together cover a timespan ranging from the earliest written attestations up to the end of the thirteenth century. Even for these long-standing projects, new digitalisation techniques are pointing the way toward profound, necessary, and fruitful changes that affect the workflow, the management of the vast amount of material, and the global accessibility of the results. After considering various proposals, a number of solutions have been adopted for each of these areas: some are already in operation, while others are currently being implemented. What follows is the report of three staff members.*

Keywords: lexicography; database; research data; DSL (domain-specific language); library.

1. Presentation of the Mittellateinisches Wörterbuch (MLW)

General presentation

The *Mittellateinisches Wörterbuch* forms part of the UAI's project *Dictionnaire(s) du Latin médiéval*, which aims to replace the seventeenth-century *DuCange (Glossarium Mediae et Infimae Latinitatis)*. This large-scale project is built on two pillars: the 'universal' dictionary *Novum Glossarium Mediae Latinitatis*, based on the extensive, universal corpus of medieval Latin texts dating from 800-1200, and around fifteen national dictionaries, each drawing on sources relevant to their linguistic regions and focusing on different time spans. Another part of the project is the journal called *Archivum Latinitatis Medii Aevi*. While some national dictionaries have already been completed, others remain in progress. The dictionaries' coverage of the alphabet remains largely disparate.

The Mittellateinisches Wörterbuch forms part of the German Akademienprogramm. It is compiled in Munich at the *Bayerische Akademie der Wissenschaften*, in cooperation with the *Österreichische Akademie der Wissenschaften* and the *Schweizerische Akademie der Geistes- und Sozialwissenschaften*. Among all the national dictionaries produced under the aegis of the UAI, the MLW is the most comprehensive dictionary. It is a complete dictionary, what we call a *Globalwörterbuch* in German: It covers all words attested during the set period and the defined territory, together with all their meanings. For meanings already well attested in the *Thesaurus Linguae Latinae*, the MLW offers a restricted selection of quoted occurrences, while medieval Latin neologisms are treated in their entirety.

Like all UAI dictionaries, the MLW focuses on semantics: The analysis of nuances of meanings dominates, but we also note, more or less exhaustively, the phraseological particularities of the lemma and offer specialized indications on the subject in question.

Text corpus (temporal and geographical limits)

The MLW covers a period of roughly eight centuries: it begins in the 6th century (where the *Thesaurus Linguae Latinae* ends) and it ends with the death of Albert the Great in 1280. Albert's works have been included in the corpus to provide an overview of the beginnings of medieval studies on Aristotle, as well as scholastic terminology. The corpus encompasses texts from all literary genres: historiography, poetry, technical treatises, legal documents, and so on. The geographical boundaries of the regions from which these texts originate correspond, more or less, to the current German-speaking regions, i.e. Germany, Austria and German-speaking Switzerland. The corpus is supplemented by further historical sources, which are essential to an understanding of German history, and by a selection of technical treatises, which exceeds the defined geographical limits and provides an overview of technical terminology, such as mathematical, alchemical, medical, legal, etc.

Scope of working materials; records

Our corpus of texts consists of around 4,000 titles of varying lengths, from which approximately one point five million records on paper slips have been generated. These slips form the basis of our work.

Actual and future plans:

In addition, with the start of the letter 'S', we have changed our data entry system from *Word* to an XML-based data entry system. This provides us not only with the usual printed version, but also with the data required for the new digital version.

Digitisation

Digitisation of the MLW

Digitisation of the already published volumes is currently underway. To this end, the articles from letters 'A' to 'E' have been digitised and prepared for presentation on the network. The work is being carried out by the *Kompetenzzentrum* at the *University of Trier*. The MLW can now be consulted on their platform (http://woerterbuchnetz.de/), alongside other dictionaries, online. After the letters A-E, it is planned to put the rest of the MLW already in print online, but with a moving wall of five fascicles. Letters "F-K" should follow in the beginning of 2026.

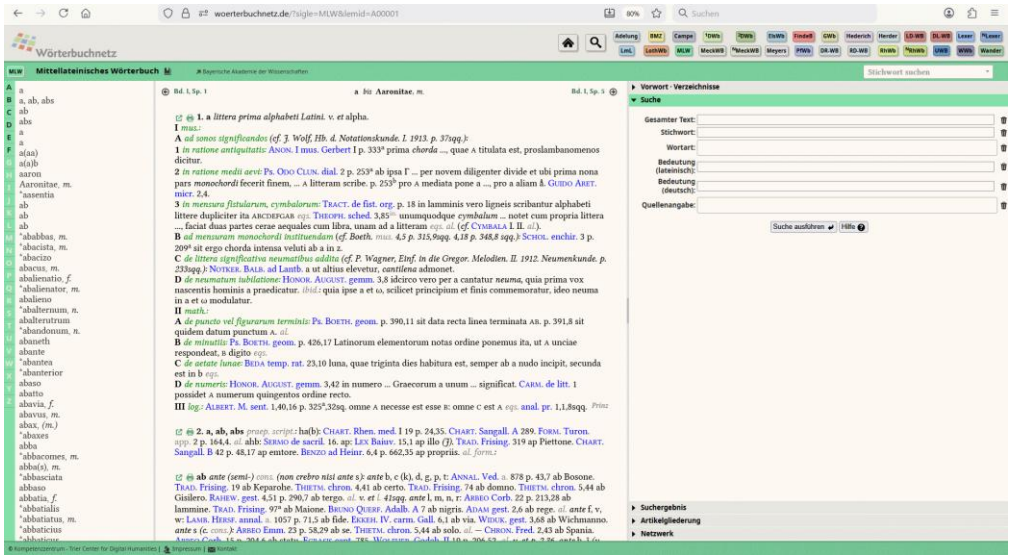

Fig. 1 – https://woerterbuchnetz.de/?sigle=MLW

The *Woerterbuchnetz* portal contains encyclopaedias, dictionaries of linguistic history, German regional dialect dictionaries and specialised dictionaries focusing on the German-speaking world. In addition to the MLW, there is a second Latin dictionary from the Middle Ages, the *Lexicon musicum Latinum medii aevi*.

The structural presentation of the MLW follows the example of other projects represented on the *Woerterbuchnetz*. A letter bar, a list of lemmata, articles, searches, the article structure, as well as access to the index of authors *("Quellenverzeichnis")* and the directory of abbreviated references *("Literaturverzeichnis")* can be found.

The following search options are available:

Search the entire text, *("Gesamttext")* search by lemma *("Stichwort")*, by part of speech *("Wortart")*, by German or Latin meanings *("Bedeutung")*, and by cited sources *("Quellenangabe")*.

As an extra service, searchable pdfs of the printed fascicles are published online on the MLW homepage (mlw.badw.de/mlw-digital.html), likewise following the five-fascicle moving wall.

Digitisation of the paper slips

Another current project is the digitisation of our paper slips. This will not only serve as a long-term archive, but helps us to accelerate the workflow of our current work, too. Parts of the digitised paper slips will be made available to the public in the future.

Digitisation of editions

In addition to our paper slip records, we use digital databases to write our articles: a large number of the sources quoted by MLW can be consulted online in their entirety. In addition, with the help of assistants, we have digitised all our other sources and established a full text, enabling us to carry out digital searches, which are very useful for our day-to-day work; in any case, as many of these editions are subject to copyright, they are currently reserved for the internal use of the MLW only.

Database

All these elements are connected in a database, which has become an indispensable working tool for us over the last few years. We are currently working on improving our database, in which the various digitised documents, like paper slips, editions, and information on authors, are being linked together. Last year, the database underwent a system update to enhance its stability and prepare for its future open-access functionality. Over the next two years, a version is planned that makes selected data freely available to registered researchers.

2. The Thesaurus linguae Latinae towards the Future without Renouncing the Past

The work of the *Thesaurus linguae Latinae* consists in writing the "biographies" of words by examining all their attestations, from the earliest Latin texts down to Isidore of Seville, that is, up to the beginning of the seventh century. Every type of text is included in the investigation of a lemma: alongside so-called literary texts, the *Thesaurus* also takes into account medical and legal writings, inscriptions, and translations of the Bible from Greek and Hebrew, among others. For this research, the lexicographer relies on the vast body of material assembled at the end of the nineteenth century precisely for the purpose of producing the first comprehensive dictionary of the Latin language. This material consists of roughly ten million slips (*Zettel*),

Fig. 2 – The 'Zettelarchiv' des Thesaurus linguae Latinae

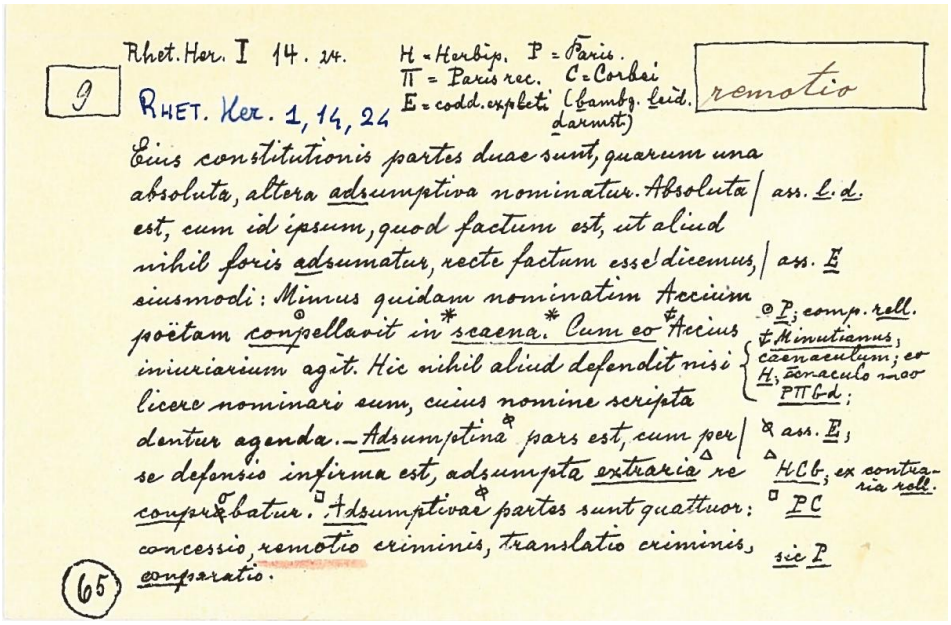

Fig. 3 – A 'Zettel' from the remotio-Material

each containing a passage in which the word in question occurs, arranged chronologically in one or more boxes depending on the number of attestations.

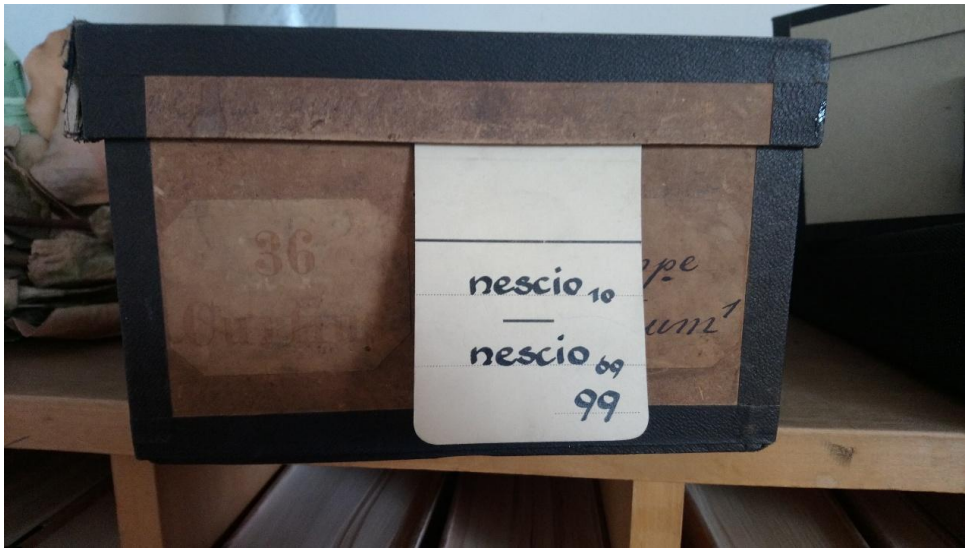

Fig. 4 – A 'Zettelkästchen' from the nescio-Material

This corpus represents a significant advantage over modern digital databases, for it is already lemmatized.

Another essential tool of the *Thesaurus* is the index of texts and authors. On the one hand, it lists the citation forms (*Zitierweise*) used in the *Thesaurus*, which have become the standard ones for most scholars, specialized journals, and publishing houses. On the other hand, it indicates what the *Thesaurus* staff – responsible for keeping track of new editions and assessing their scholarly reliability – regard as the best editions of ancient works. The index is closely connected with the *Thesaurus* library, itself organized chronologically and comprising all Latin authors and works, together with the relevant secondary literature (commentaries, etc.).

Once a lexicographer is assigned a lemma, they receive the corresponding material – the *Zettel* – and begin reading each passage with close attention to all lexicological features of the word, to its uses and their differences. It is precisely through these differences that the lexicographer constructs a layout of senses (*Gliederung*): a system of hierarchically organized levels and sublevels that interact with one another and that, starting from the most evident distinctions (e.g. *proprie* vs. *translate*, or *usu transitivo* vs. *usu intransitivo* and so on), moves toward subtler nuances not immediately visible at first glance. The *Gliederung* thus becomes a kind of genealogical tree of meanings, which not only enables users to identify the specific sense relevant to their inquiry, but also to understand how the word developed that meaning in a given context.

Because of the various challenges that the compilation of the *Thesaurus* has always posed, the advent of new digital technologies was regarded from the outset as a potential aid. As a result, the earliest CD-ROMs – such as the PHI (Packard Humanities Institute) disk – soon became part of the *Thesaurus* workflow. In 2002, in collaboration with the publishing house Saur (which had just acquired the historic Teubner), the first CD-ROM of the *Thesaurus* was released; it contained all printed material up to the letter *P*. This was an initial step – technically still immature, to be sure – yet necessary to bring the TLL into the digital age.

In 2008, the publishing house – by then De Gruyter – launched the online version of the *Thesaurus linguae Latinae*, which included all dictionary articles published from 1900 to 2008, as well as the special volumes of proper names and the *Index*.

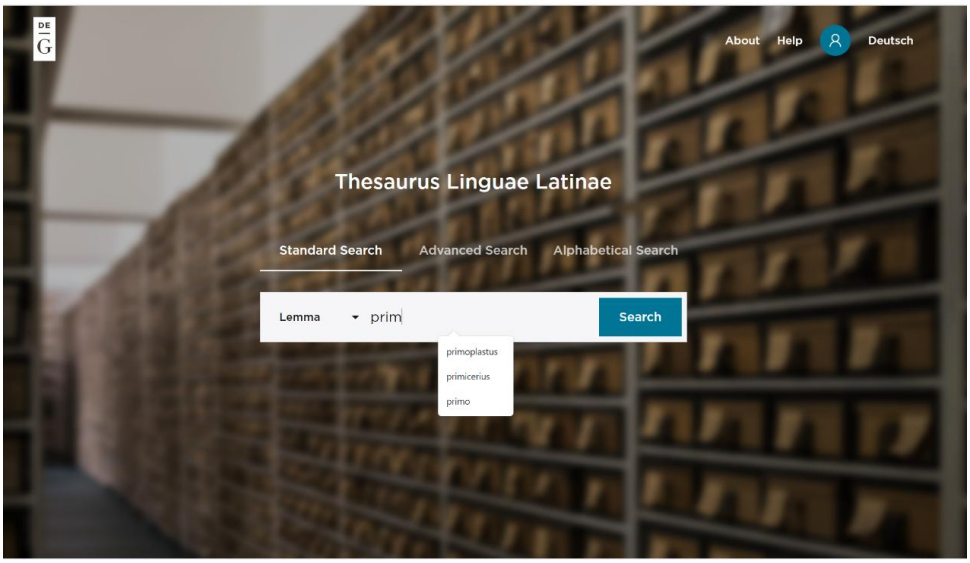

Fig. 5 – The Thesaurus database by De Gruyter

The digital database employs an XML-based encoding to represent, among other things, the hierarchical structure of entries, textual segments such as citations, and the markup of editorial abbreviations typical of the TLL. The major limitation of this version is that it requires a paid license and is therefore accessible only through an institutional or library subscription.

To ensure that the results of scholarly research be freely accessible, a few years ago – thanks above all to the efforts of Johannes Ramminger and the IT department – a freely available Open Access version of the *Thesaurus* was released on the server of the Bavarian Academy of Sciences.

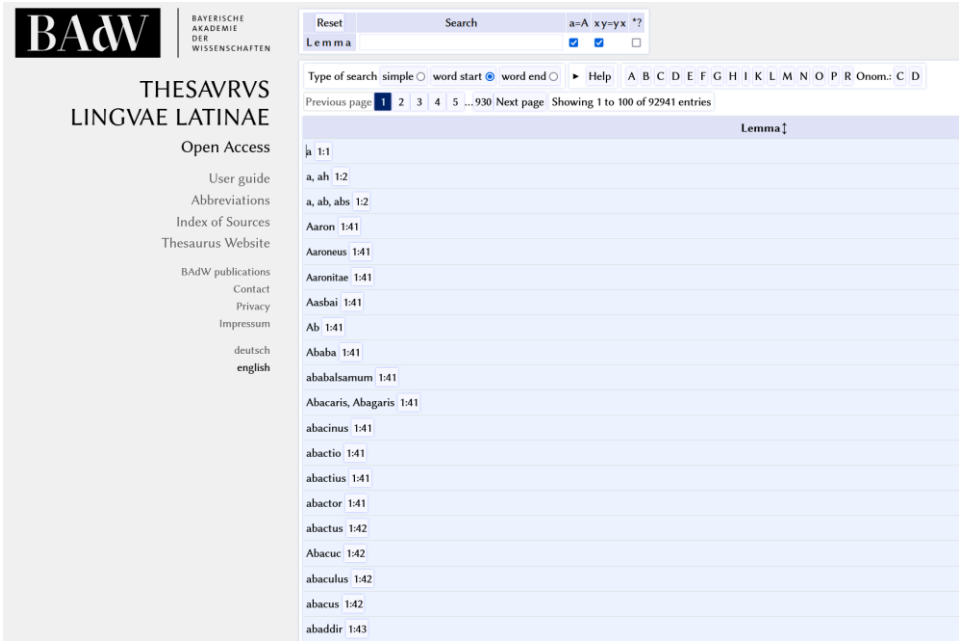

Fig. 6 – The open-access PDF version – https://thesaurus.badw.de/tll-digital/tll-open-access.html

This version offers various functionalities and also includes the *Index* in tabular form. Its only constraint is a three-year moving wall: the most recent articles become accessible only three years after publication.

Even this solution, though extremely helpful for many scholars wishing to consult the *Thesaurus* from any location, must be regarded only as one stage in the broader process of digitalizing the project. Ensuring that the *Thesaurus* maintains its status and its mission as an indispensable tool for Latin studies requires making its results even more readily accessible. To achieve this goal, several challenges must be addressed: above all, the digitalization of the enormous archival material, followed by the digitalization of the index; the retro-digitalization of the already printed articles; and, finally, the creation of a new input system that would produce *Thesaurus* entries encoded in XML from the very moment they leave the Munich workshop.

As far as retro-digitalization is concerned, in July 2024 a five-year agreement was signed between Sapienza University of Rome and the Bayerische Akademie der Wissenschaften to launch an experimental project aimed at producing a new digitalization of the *Thesaurus linguae Latinae* (ThlL). The project is directed by Prof. Michela Rosellini and Dr. Elena Spangenberg Yanes, with the collaboration of Andrea Consalvi and Daniele Fusi.

Starting from a preliminary analysis of the functionalities offered by the current digital version published by De Gruyter, this initiative aims to develop an innovative markup model that not only reproduces the typographical layout of the entries (their formatting), but also introduces a semantically structured encoding articulated on three distinct layers: (1) the structure of the entry – covering the lemma section, the *Kopf* with its various rubrics, the *Gliederung* with its hierarchical levels, and the stylistic appendix with its subdivisions; (2) the citations, which form the core of the lexicographical text; (3) the linguistic plane, with

particular attention to elements that can be automatically recognized, such as literal vs. figurative uses, or abbreviations marking technical terminology.

This advanced markup will significantly expand the hermeneutic and research possibilities, making it possible to filter the content according to parameters such as author, work, period, genre, or even the position of the citation within the *Kopf* or the *Gliederung*. Users will be able to combine these filters freely – for example, to extract all adverbs used within a given period and within a specific literary genre, to retrieve all grammatical technical terminology, or to observe the degree of terminological overlap among two or more authors.

With these developments – and with those discussed by Nora Götze in her contribution – the *Thesaurus linguae Latinae* will undoubtedly be able to maintain the fundamental role it has always played in the field of Latin studies.

3. Building Blocks of Digital Lexicography

Overview

At MLW and TLL, we are working on digital tools in the following areas, which have been fundamental to lexicography even before the digital age:

- boxes of paper slips with lemmatized material
- indices of the lexicographical sources
- a library of texts, both within and out of copyright restrictions
- preparation and production of dictionary articles
- the publication of research data

They might seem straight-forward at first glance, but are actually each a world of their own, both for the complex interconnection between them and the additional requirement of them being useful for the interested public.

A digital Zettelkasten

As seen, most lexicographical projects work with "Zettel". There might be millions of them, organized in boxes of about 1400 paper slips. In MLW, a prototype for a digital database of paper slips was introduced in 2020 and in active use ever since. At this point, their paper slips have all been scanned and categorized. In TLL, scans are currently being made from microfilms taken in the 1960es. There will be well beyond 9.000.000 paper slips to be digitized in the upcoming years.

The paper slips are a useful tool and necessary for work outside the archive, and they also have to be published as part of the open access initiative for research data at the BAdW.

A paper slip contains similar information across projects. There is, generally,

- a lemma
- an author/work abbreviation and citation
- a quote

Some paper slips might deviate from this outline, but they still follow a general rule of referring to a lemma and the work that it has appeared in.

We tried to arrange this data in a digital database, so that people can sort and search for it in a digital *Zettelkasten*, which is still in development but will be finished soon.

Paper slips can now be enriched with personal commentary, which has been done offline directly on the paper slips. It will also be possible to create native digital paper slips, foregoing the actual paper slips in the future.

The *Zettelkasten* is based on three pillars: the scanned paper slips, the list of lemmas to attach to these paper slips (more or less a mix of the official lemmalist and those words that were not included in the dictionary), and similarly, an index of sources including both reference editions and further material.

In the *Zettelkasten*, each paper slip can be connected to a lemma and a number of references. Once connected, it is possible to find paper slips by their lemma or the works that are being referred to, and to navigate from paper slips to works and reference editions. Staff working on articles can organize paper slips into piles and annotate them to their personal preference. In the future, a simple export mechanism will deliver a basis for the written article.

A digital index of sources

A digital index already exists for both the MLW and the TLL, containing URLs to sources as well. However, both are mostly structured as to emulate the printed indices for these dictionaries. In order to make them useful for the purpose of the *Zettelkasten* and, additionally, to provide further ways of access for a wider scientific community, it is necessary to categorize this information in a machine-readable way.

So far, information is implied and mixed up in the indices. Examples for this are:

- dating criteria and descriptions are collected at the same level for works and authors
- abbreviations for authors and work are not separated unambiguously
- norm-data are sprinkled throughout
- editions are collected as a group for each work, sometimes for an author, some with URLs

The most important part of digitizing the indices is therefore to organize the data precisely, which we are currently in the process of doing.

A digital library

A digital library would, perhaps, be one of the most rewarding additions to the tools. Of course, digital libraries do exist, but so do copyright laws. That is why, from a certain point in time, the reference editions are not yet public, or free. Under copyright law, we are allowed to create and distribute among the projects digital scans of the editions that they hold in their respective libraries. Currently, this manifests in an unwieldy internal PDF library, and a lot of OCR that has not been proofread, yet. At MLW, attempts were made to produce them in a viewer with some search capability, at least.

What would be better, though, is a fully-fledged reference library to help with and provide quotes and citation information for references. The Perseus SCAIFE viewer (https://scaife.perseus.org/) seems like a great base for such a system, although it would need

to be able to ingest PDF and show it as part of the text view to make dealing with faulty OCR easier. Another desideratum would be the possibility to annotate pages, which has been done by generations of scientists in the margins of the library copies until now.

Article production

At MLW, article production follows an intuitive system that has been developed between the project itself and the digital humanities department of BAdW. Basically, they have devised a machine-readable, very flexible MLW grammar. This gives them full control over their data output: an MLW article can go into a print facsimile just as soon as into the XML format of the *Wörterbuchnetz Trier* (https://woerterbuchnetz.de/) or, in fact, any hypothetical publishing website.

MLW authors write their articles in a custom-built MLW language. The files are then processed via the DH Parser (https://github.com/jecki/DHParser), which was developed in-house at the BadW to create an XML file that can be further processed for use in the Wörterbuchnetz, or transformed into a PDF file, or HTML to be published, should they so choose.

Fair data

We are required to follow "fair principles" of publishing research data. This means that accompanying the printed edition of the dictionaries, the project's research data should be as easily accessible and interconnectable as possible. We are currently looking into exchange formats and ways to make sure the data follows certain standards: For example, the open access documents accessible by lemma id will use the same id as in the digital *Zettelkasten* for that same lemma.