# CLARINO+ Optimization of Wittgenstein Research Tools

**Alois Pichler**
Wittgenstein Archives at the University of Bergen
Bergen, Norway
alois.pichler@uib.no

## Abstract

The Wittgenstein Archives at the University of Bergen (WAB) offer specialized tools for research access to its Wittgenstein resources which however are in need for an upgrade to better serve user requirements. The paper discusses this need along some selected exemplary features of two such tools: Interactive Dynamic Presentation (IDP) of Wittgenstein's philosophical Nachlass and Semantic Faceted Search and Browsing (SFB) of Wittgenstein metadata. The tasks of extending and better adapting these two tools to user requirements shall be carried out within the Norwegian CLARINO+ project.

## 1 Data and metadata for Wittgenstein research

During his lifetime, the Austrian-British philosopher Ludwig Wittgenstein (1889-1951) published only one philosophical book, the *Logisch-philosophische Abhandlung* / *Tractatus logico-philosophicus* (1st ed. 1921/22), and a *Dictionary for Elementary Schools* (1st ed. 1926). However, on his death in 1951, he left behind a significant 20,000 page corpus of unpublished philosophical notebooks, manuscripts, typescripts and dictations. This oeuvre, called "Wittgenstein's Nachlass" or "the Wittgenstein papers" (von Wright, 1969), was brought to the wider public through posthumous book publications such as *Philosophical Investigations* (1st ed. 1953) and *Culture and Value* (1st ed. 1977).

The practice of bringing the Nachlass to modern readers through digital editing, hereby creating new access and research possibilities, reached its first milestone in 1998 with Vol. 1 of the Bergen CD-ROM edition *Wittgenstein's Nachlass: The Bergen Electronic Edition* (Wittgenstein, 2000), edited by the Wittgenstein Archives at the University of Bergen (WAB, http://wab.uib.no/). Since its establishment in 1990, WAB has worked towards providing digital data and metadata for conducting Wittgenstein Nachlass research (Huitfeldt, 2006). This includes the creation of machine-readable transcriptions of the Nachlass with specialized markup. These transcriptions are today accessible as HTML outputs through "interactive dynamic presentation" interfaces (IDP, see Pichler and Bruvik, 2014) on WAB's "Nachlass transcriptions" site http://wittgensteinonline.no/ (Wittgenstein, 2016-). Along with high quality Nachlass facsimiles, they are also increasingly available as HTML outputs in WAB's *Bergen Nachlass Edition* on Wittgenstein Source (Wittgenstein, 2015-). In addition, WAB is working on the implementation of semantic web methods and technology for Wittgenstein research and offers free download of a continuously growing Wittgenstein ontology (see Pichler and Zöllner-Weber, 2012) in OWL (RDF) format from its website, as well as an ontology explorer for semantic faceted search and browsing (SFB, http://wab.uib.no/sfb) of the ontology.

However, WAB's transcriptions and facsimiles of the Nachlass, its metadata for the Nachlass and Wittgenstein research more generally, as well as the tools for making all these accessible are in

need of upgrades for the research community to be able to make the most of them. In this paper I shall focus on some selected exemplary aspects of the required optimization of the IDP and SFB tools. The tasks of extending and better adapting these two tools to user requirements shall be carried out within the Norwegian CLARINO+ project.

## 2 Interactive Dynamic Presentation (IDP) of Wittgenstein's Nachlass

Digital editions offer significant advantages over print editions in that they allow dynamic and user-tailored access to the material edited. WAB's transcriptions of the Wittgenstein Nachlass contain XML TEI (P5) markup with detailed philological and semantic information about each of the Nachlass items, pages, remarks, sentences, formulas, drawings, words, letters, and characters. The IDP tool to access these transcriptions is built on XML technologies (using Xalan XSLT), HTML and PHP, and permits the user to produce HTML clean copy, "linear" outputs of the Nachlass as well as "diplomatic" outputs recording all deletions, insertions, overwritings etc.[2]

At first, most users simply want to access, read and search the Nachlass through these two primary presentation formats. But they soon discover that they still need other types of versions: for example, a version that in contrast to the diplomatic output omits those deletions where the parts deleted don't fit syntactically into the context and thus "disturb" the reading; or a version that in contrast to the standard linear rendering still retains deleted parts (but shows that they are deleted by Wittgenstein) the inclusion of which may help understand the text; or a version which in contrast to the linear version offers full up-to-date standardization of orthography and punctuation. Yet other requirements include the possibility to filter and arrange the Nachlass corpus according to the editorial marks which Wittgenstein often assigns to his remarks, while at the same time also retaining the possibility to include or omit the marks themselves in the editorial output. This feature permits, for example, extraction of all remarks and only the remarks which are marked by Wittgenstein with a slash, or an asterisk, a backslash, etc., or a specific combination of them, and can help the user see the genetic processes behind the Nachlass or recognize thematic groups. Still another required feature is the possibility to interactively remove in the transcription of a typescript all handwritten revisions and thus to produce a version of the typescript in its purely machine-typed form. This function comes handy when one wants to compare vocabulary and concepts before and after the revision and was indirectly already asked for almost fifty years ago when A. Kenny (1976) wished for an edition of the Big Typescript "as it stood" (i.e. before Wittgenstein's revisions of it). Finally, yet other features required and crucial to Wittgenstein Nachlass research include the possibility to organize the remarks of an item chronologically rather than in their physical order (which often deviates from the chronological one), and the possibility to arrange Nachlass texts according to Wittgenstein's editorial numbers that he uses to group his remarks according to topic.

The possibility to control WAB Nachlass transcription outputs through IDP tool parameters such as the ones mentioned above is valuable to Wittgenstein research, and many of the features described are prepared for in WAB's transcriptions through specific XML TEI encoding. Chronological sorting, for example, can be implemented on the basis of the fact that WAB's transcription records contain a date for each single Nachlass remark; omission of handwritten revision in typescripts can be achieved thanks to WAB's explicit encoding of handwriting in typescripts; filtering and sorting of the Nachlass texts according to Wittgenstein's editorial numbers could be put to practice thanks to the specific encoding WAB uses for them, etc. etc. But to put all relevant encodings to work and to offer them for IDP toggling demands a substantial programming investment. While some of the features required and mentioned above already work, many do not yet work flawlessly or on all required levels. Chronological sorting, for example, currently only works on the level of single items, although it is precisely at higher levels that it may be needed the most, for example, in the chronological arrangement of Nachlass item *groups* or even the *entire* Nachlass corpus. In brief, the IDP tool currently manages to offer access to (1) only a fraction of the encoding, (2) only a fraction of combinatorial possibilities of the encoding, (3) only a fraction of the presentation, sorting and filtering possibilities, and (4) in all these three fields it is susceptible to errors due to undesired interference. It is in fact a major challenge to provide for presentation modes

[2] For a more thorough explanation of the terms "diplomatic" and "linear", see Pierazzo (2009).

that work in tandem and that can be accumulated rather than interfering negatively with each other. This factor results in limitations such as the following: with regard to (1), users cannot yet filter the transcriptions for insertions of a specific subtype; with regard to (2), users are not yet able to combine filtering of insertions with filtering of the encoding of text alternatives; and with regard to (3), it is not yet possible to render the type of insertions selected in ways other than what is set by WAB as the default for the IDP site. Moreover, with regard to (2), it is for the user currently not possible to combine a marking of Wittgenstein's text alternatives with a diplomatic rendering, or with the inclusion / exclusion of his own markers for text alternative, or with a toggling of including / excluding the alternatives discarded by him.

Users with XML programming competence may be able to respond to all such needs by directly processing and querying the XML transcription file itself. This is one of the reasons why also within the framework of the CLARINO+ project WAB will deposit its transcription corpus in the CLARINO Bergen Repository. This task includes generation of CLARIN CMDI-conformant metadata as well as designing licenses for the use of both the transcriptions and the metadata offered. Previously within the frameworks of the Cost A32 and Discovery projects, WAB has already made available XML transcription samples of 5000 Nachlass pages under the CC BY-NC 4.0 license.[3]

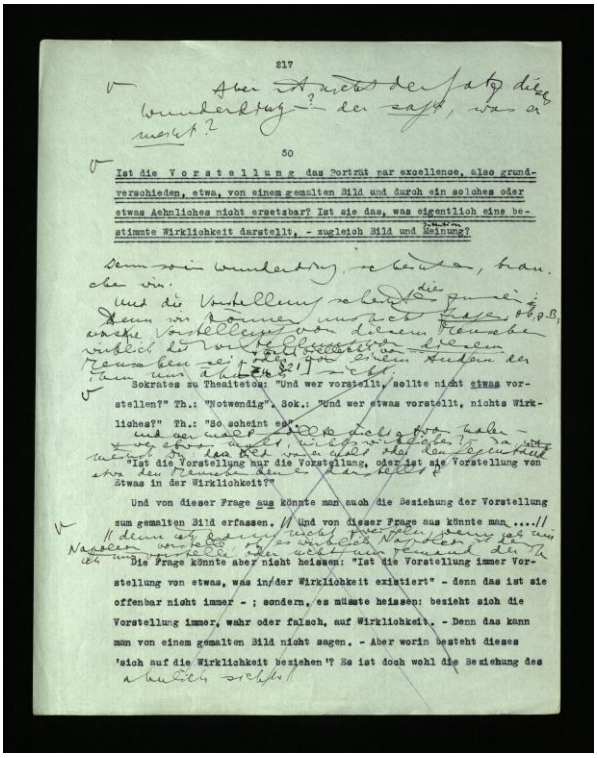

*Figure 1: Facsimile of Wittgenstein Nachlass Ts-213,217r, reproduced with the kind permission of The Master and Fellows of Trinity College Cambridge and the University of Bergen. CC BY-NC 4.0.* `http://www.wittgensteinsource.org/Ts-213,217r_f`

*The page displays (most relevant for IDP) handwritten revisions of the typescript incl. deletions, insertions, markings, variant writing and frequent use of the editorial mark "v". The page also contains (most relevant for SFB) at remark Ts-213,217r[3], "Sokrates zu Theaitetos …", a reference to Plato, as well as the internal reference "[Zu § 21]". Transcriptions of this remark are made available by WAB at* `http://www.wittgensteinsource.org/Ts-213,217r[3]_d` *(diplomatic version),* `http://www.wittgensteinsource.org/Ts-213,217r[3]_n` *(linear version) and (interactive dynamic presentation)* `http://wittgensteinonline.no/`.

## 3 Semantic Faceted Search and Browsing (SFB) of Wittgenstein metadata

WAB's reference system assigns a unique identifier to each remark in the Nachlass, called "siglum". The siglum provides a URL for each single Nachlass component and makes up the backbone of the Wittgenstein ontology and the SFB site that offers semantic faceted search and browsing of WAB's metadata for the Wittgenstein domain. The SFB tool is built on the University of Bergen Library's search infrastructure, which involves technologies such as Elasticsearch,

[3] See `http://wab.uib.no/cost-a32_xml/`; for the aforementioned two projects see `http://wab.uib.no/wab_R&D.page`. It must be noted that the entire Wittgenstein Nachlass is made available by WAB as open access in the sense of *free*, but only the mentioned 5000 pages are currently open access also in the sense of *libre* open access (for the distinction see Suber, 2003). For the licenses for all WAB resources currently offered to the public, see Pichler (2019).

Apache Jena and Angular framework.[4] Today, SFB already permits search and browsing of Nachlass remarks along a number of facets, incl. reference to a person, reference to a work, its dating and its relation to "published works", and displays the resulting remark hit along with a link to the corresponding facsimile in Wittgenstein (2015-). However, while much more metadata are recorded in the transcription or by stand-off markup, they need first to be modeled and ingested into the tool. Examples are information about a remark's genetic path(s), its place of origin in the Nachlass corpus, references to places, events and other named entities, similarity to other remarks (see Ullrich, 2019), adherence to text type and genre (philosophical remark, preface, motto, dedication, instruction, aphorism, diary entry, autobiographical remark, mathematical-logical notation, graphic etc.), adherence to Nachlass group (notebook, loose sheet, "Zettel", ledger, typescript, dictation etc.), work status (first draft, elaborated version, final work etc.), script type (short hand, secret code etc.), the language the remark is written in, the writing material (pencil, ink etc.), research literature referring to it, and other.

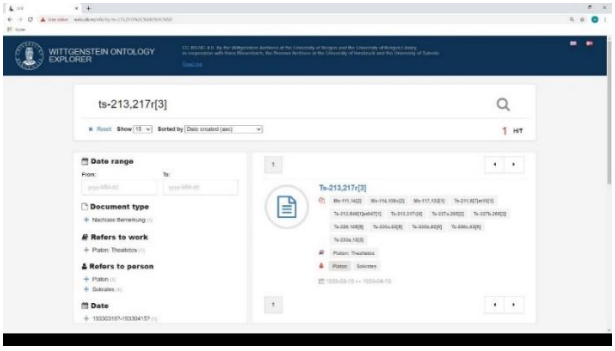


*Figure 2: Screenshot of the SFB hit for Wittgenstein Nachlass remark Ts-213,217r[3], available at* `http://wab.uib.no/sfb/?q=ts-213,217r%5C%5B3%5C%5D`

An important contribution to the SFB service will be the inclusion of a digital Wittgenstein lexicon (see Röhrer, 2019), as an outcome of the cooperation between WAB and the Centrum für Informations- und Sprachverarbeitung (CIS) at the Ludwig Maximilians Universität München on the search tool WiTTFind (`http://wittfind.cis.lmu.de/`) – WAB contributing its facsimiles and encoded XML transcriptions of the Wittgenstein Nachlass as well as XSLT stylesheets for their processing, and CIS providing programming and computational linguistics personnel resources as well as a grammatically encoded digital lexicon of the German language (see Hadersbeck, Pichler et al., 2016). WiTTFind offers lemmatized online text search access to the entire Nachlass, displays each sentence containing any grammatical form of the word searched for within the context of the larger remark, and additionally highlights the hit in the corresponding facsimile of the remark. WiTTFind continues WAB's siglum reference system for the Nachlass even down to sentence level. Recently, the cooperation project has also embarked on a word tokenizer based on WAB's XML transcriptions of the Nachlass. Implementing WiTTfind in the SFB tool will permit simultaneous and combined SFB of *both metadata and text data*. Researchers interested in the genesis of Wittgenstein's philosophy, for example, may want to know when Wittgenstein started to replace the expression "calculus" with the expression "game", and whether this development can be linked to any other development, e.g. increased reference to works of others, other changes in vocabulary, developments in letter correspondence, meetings and discussions with friends and colleagues, etc. An integration of the SFB and WiTTFind tools will bring us closer to seeing all the connections between the Nachlass' remarks and contents. WiTTFind is a fine example of the added value created by making one's data available for research and reuse by others, and CLARINO+ is a fine example of capitalizing this value further by implementing its outcomes into the CLARIN confederation of language and text resources.

CLARINO+ will both integrate the WiTTFind lexicon into CLARIN and improve the SFB tool itself. Outstanding tasks include correcting errors and deficiencies in the overall browsing and combinatorial setup and adding and organizing facets still lacking; one example is chronological sorting of a remark's variants, which currently are only displayable in alphanumeric order (for an

[4] See `http://marcus.uib.no`, `https://www.elastic.co`, `https://jena.apache.org/` and `https://angular.io/`.

example see Figure 2). The upgrade will require improvements of the user interface, including addition of display labels for creation dates along with search results. A highly desired addendum is the possibility to view the remark hit resulting from one's searching and browsing along with a linear or diplomatic transcription of the remark; currently only the remark's siglum along with a link to the corresponding facsimile is displayed.

## 4 Conclusion

Although at present WAB, along with its IDP and SFB tools, already enjoys a large number of international users[5], it is only when deficiencies such as the ones described above are corrected and further requirements and desiderata fulfilled, that researchers will be able to take full advantage of WAB's resources. Only then will users be equipped to fully exploit the multifaceted interrelations between and within Wittgenstein data and metadata provided by WAB for the community's research questions. At the same time, it is also then that the deep issues about the relation between on the one hand the contents and forms of Wittgenstein's philosophy and work, and on the other hand their interpretation and application, can properly begin to play out in sufficiently complex formats via interactive digital media.

## References


Max Hadersbeck, Alois Pichler, Daniel Bruder, Stefan Schweter. 2016. *New (Re)Search Possibilities for Wittgenstein's Nachlass II: Advanced Search, Navigation and Feedback with the FinderApp WiTTFind*, in: Contributions of the Austrian Ludwig Wittgenstein Society, 90–93, Kirchb. A. W.

Claus Huitfeldt. 2006. *Philosophy Case Study*, in: Electronic Textual Editing, Modern Language Association of America, 181-196.

Antony Kenny. 1976. *From the Big Typescript to the Philosophical Grammar*, in: J. Hintikka (ed.), Essays on Wittgenstein in Honour of G. H. Von Wright, Acta Philosophica Fennica, 28, 41–53.

Alois Pichler and Amelie Zöllner-Weber. 2013. *Sharing and debating Wittgenstein by using an ontology*, Literary and Linguistic Computing, 28 (4), 700–707.

Alois Pichler and Tone Merete Bruvik. 2014. *Digital Critical Editing: Separating Encoding from Presentation*, in: D. Apollon, C. Bélisle, Ph. Régnier (eds.), Digital Critical Editions, 179–199, Urbana Champaign.

Alois Pichler. 2019. *A brief update on editions offered by the Wittgenstein Archives at the University of Bergen and licences for their use (as of June 2018)*, in: Wittgenstein-Studien, 10(1), 139–146.

Elena Pierazzo. 2009. *Digital genetic editions: the encoding of time in manuscript transcription*, in: M. Deegan, K. Sutherland (eds.), Text Editing, Print and the Digital World, 169–186, Famham.

Ines Röhrer. 2019. *Lexikon, Syntax und Semantik – computerlinguistische Untersuchungen zum Nachlass Ludwig Wittgensteins*, Master's thesis at LMU München, Munich.

Peter Suber. 2003. *Removing the Barriers to Research: An Introduction to Open Access for Librarians*, in: College & Research Libraries News, 64, 92–94, 113 [unabridged online version at http://legacy.earlham.edu/~peters/writing/acrl.htm].

Sabine Ullrich. 2019. *Boosting Performance of a Similarity Detection System using State of the Art Clustering Algorithms*, Master's thesis at LMU München, Munich.

Ludwig Wittgenstein. 2000. *Wittgenstein's Nachlass: The Bergen Electronic Edition*, ed. by the Wittgenstein Archives at the University of Bergen under the direction of Claus Huitfeldt, Oxford.

Ludwig Wittgenstein. 2015–. *Wittgenstein Source Bergen Nachlass Edition*, ed. by the Wittgenstein Archives at the University of Bergen under the direction of Alois Pichler, in: Wittgenstein Source (2009–) [wittgensteinsource.org], Bergen.

Ludwig Wittgenstein. 2016–. Interactive Dynamic Presentation (IDP) of Ludwig Wittgenstein's philosophical Nachlass [http://wittgensteinonline.no/], ed. by the Wittgenstein Archives at the University of Bergen under the direction of Alois Pichler, Bergen.

G.H. von Wright. 1969. *The Wittgenstein papers*, The Philosophical Review 78(4), 483–503.


[5] Google Analytics lists for `http://wittgensteinonline.no/` more than 4 800 and for `http://wab.uib.no/sfb/` more than 1 400 users since 2017. I would like to thank my Bergen colleagues Nivedita Gangopadhyay, Øyvind Gjesdal and Hemed Al Ruwehy for comments on a draft of this paper.