CLSINFRA COMPUTATIONAL LITERARY STUDIES INFRASTRUCTURE

D3.1

Baseline Methodological User Needs Analysis

Authors: Christof Schöch, Evgeniia Fileva, Julia Dudar

Date: February 28, 2022

This project has received funding from the European Union’s Horizon 2020 research and innovation programme under grant agreement No 101004984

Project Acronym: CLS INFRA

Project Full Title: Computational Literary Studies Infrastructure

Grant Agreement No.: 101004984

Deliverable/Document Information

Deliverable No.: D3.1

Deliverable Title: Baseline Methodological user needs analysis

Authors: Christof Schöch, Evgeniia Fileva, Julia Dudar

Dissemination Level: PUBLIC

Document History

Version/Date	Changes/Approval	Author/Approved by
v1.0.0, Feb 28, 2022	Additional analysis.	Schöch, Fileva, Dudar
v0.9.0, Feb 24, 2022	Initial full version.	Schöch, Fileva, Dudar

Table of Contents

1. Introduction

The findings reported here have been obtained in the framework of CLS INFRA's Work Package 3 concerned with "Methodological Considerations of Computational Literary Studies" (WP3). The overarching objective of WP3 is to identify, document and show-case current shared practices in CLS research. This objective supports several purposes, among them guiding infrastructure development, defining training opportunities, and consolidating the CLS community. Specifically, one can deduct infrastructure requirements from such findings and feed them to other work packages within the project, in order to ensure that decisions taken when designing a research infrastructure for the CLS community are informed by these requirements.[1] Also, such a documentation of current shared practices can help design a useful training programme both for scholars who are newly entering the field and for more experienced researchers. In addition, it can help consolidate the CLS community by making shared research practices visible. Within WP3, the first step to identify, document and show-case best practices in CLS research is to capture the current state of the art in terms of widespread research practices in CLS. Task 3.1 on the "Baseline Methodological User Needs Analysis" is devoted to this aim.

We are not the first, of course, to engage in such an attempt to document the practices in the Digital Humanities. It has been customary so far, however, to look at the Digital Humanities as a whole, rather than at a specific subfield, such as CLS, within the wider Digital Humanities. Some recent examples are relevant in this context. Launched in 2014 in the context of DARIAH-DE, the Taxonomy of Digital Research Methods in the Humanities (TaDiRAH) is an effort to describe and structure digital research methods relevant to the Humanities.[2] TaDiRAH's focus is on providing a multilingual controlled vocabulary that is now being reused in a wide number of settings, for example as keywords for conference papers or for describing Digital Humanities courses (see Borek et al. 2016, 2021). A recent, empirical study based on a large corpus of journal articles focused more broadly on the discursive profile and topic-based disciplinary network in which the Digital Humanities as a field can be placed (Luhmann & Burghardt 2019). Another recent study aimed to map the tool usage in the field of Digital Humanities, based on the proceedings of five issues of the annual Digital Humanities conference. The results show that general-purpose tools like the programming language Python, or communication platforms like Twitter, are mentioned more frequently than even the most popular tool in the strict sense,

[1] Two good starting points for thinking about the relationship between research practices and infrastructure requirements are Moulin 2011, Svensson 2016 (notably chapter 4) and Kitchin 2021. Recent musings on infrastructure and Digital Humanities can also be found in Pawlicka-Deger (2021-2022).

[2] See: https://vocabs.dariah.eu/tadirah/en/.

which happens to be Gephi (for details, see Bardot et al. 2019, and the interactive visualizations linked there).

With a more decisive focus on CLS than most if not all previous studies, and in order to document shared research practices and provide an empirical basis for the definition of training needs and infrastructural requirements, we have hence started by building a corpus of publications that can be identified as belonging to the field of CLS and that have appeared in the timeframe 2010-2021. As a useful proxy for the currently most widespread research practices in CLS, we have then identified the frequency with which a wide range of (a) tools and software, (b) data formats, and (c) methods of analysis are mentioned in this corpus. We have identified the most widely mentioned formats, tools and methods, the development of their mentions over time, and attempted to explain the quantitative results. The findings from this analysis are documented in the present report.

The key outcomes of our study, which was conducted from May 2021 to February 2022, consist of the following elements:

- A corpus of research articles, documented by several metadata tables that includes key information on the collected publications, and containing publications marked as belonging to the field of Digital Humanities more generally, or to CLS more specifically (see chapter 2).
- A dataset and a collection of visualizations that provide information about the frequency and distribution of mentions of tools, formats and methods in the corpus mentioned above (see chapter 3).
- The present report that outlines the composition of the corpus used, explains the methodological steps undertaken, summarizes the key findings based on the data and derives conclusions from these findings.

The report consists of 5 sections. The present introduction provides basic information about the report in the context of the objectives of CLS INFRA and WP3. Section 2 is devoted to a description of the corpus of research articles and the process of collecting the data. Section 3 is devoted to an exposition of methodological concerns and the approach taken in the study, notably regarding the issue of how to best define the scope of terms relevant to this study. Section 4 summarizes the main results obtained in the study, presenting the frequency of mentions and their temporal evolution for tools, data formats and methods. Our report closes with a reflection on the conclusions one may be able to draw from the results as well as on possible future research.

2. Corpus

Our corpus consists of more than 5000 journal articles and conference papers from venues and publications that stand in a close relationship with Digital Humanities and are written in a number of different languages.

- With respect to conference papers, the articles and abstracts from the following events are included: the annual Digital Humanities Conference (organized by the Alliance of Digital Humanities Organizations, ADHO), the annual conference of the Associazione per l'Informatica Umanistica e la Cultura Digitale (AIUCD), Computational Humanities Research (CHR), DHBenelux, Digital Humanities im deutschsprachigen Raum (DHd), Digital Humanities in the Nordic and Baltic Countries (DHN/DHNB), Humanidades Digitales Hispánicas (hdh).
- Regarding scientific journals, articles from the following journals are included: Journal of Cultural Analytics (CA), Digital Studies / Le champ numérique (DS/CN), Humanités numériques (HN), Journal of Digital Humanities (JDH), Zeitschrift für digitale Geisteswissenschaften (ZfdG), Digital Humanities Quarterly (DHQ), Literary and Linguistic Computing / Digital Scholarship in the Humanities (LLC/DSH), and Language Resources and Evaluation (LREC).

The choice of conferences and scientific journals was mostly determined by the following factors: their wider prevalence within the DH community, their timeframe of activity, and their availability in open access. The time frame was taken from 2010 to June 2021.[3] Thus, we ensured that our research corpus is relevant and up-to-date. Note that not all conferences and papers included in the corpus have been published in open access with a suitable license; as a consequence, the corpus also contains closed access publications. This is a limiting factor that, unfortunately, must be taken into account when considering further work with the corpus.[4]

[3] There are only three publications from 2010 in the corpus, which were irrelevant for our study, so they were not included in the CLS data set. For this reason, further charts show results for the time period from 2011 to 2021.

[4] As far as copyright regulations allow, this corpus is made available for others to reuse. In addition, a set of tables with corpus metadata and raw data resulting from our analyses is made available for inspection and/or reuse. All data can be found on Zenodo, DOI: https://doi.org/10.5281/zenodo.6281920.

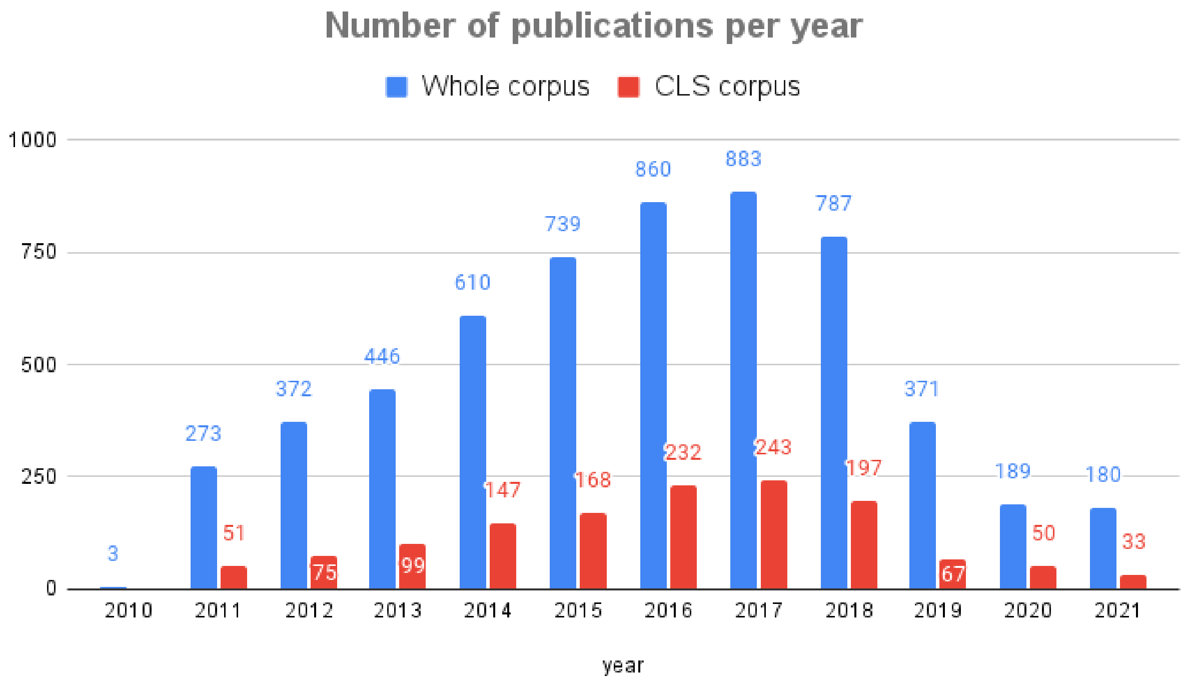

Figure 1. Number of publications in each year in two types of corpus. The exact amount of research articles in the initial corpus and in the CLS data set is shown.

The corpus building process was as follows: Scientific journals and conferences were selected according to the time frame 2010 to June 2021. Figure 1 demonstrates how the publications are distributed during that time frame and shows the number of articles per year (1) for the entire corpus and (2) for the subset of publications in the corpus that can be assigned to CLS (more on this below). Then we downloaded all available articles and papers. The publications were available in several formats: as XML, HTML, PDF, plain text (either single files or complete books of abstracts) and EPUB. After the materials were obtained, it was necessary to transform them into a plain text format. It was the most appropriate format for the purpose of our study. Converting XML to TXT was relatively easy and was done using a single Python script. PDF files were converted to TXT with the ABBYY FineReader tool. The Books of Abstracts had to be manually split into separate texts before converting them to TXT .

The resulting, initial dataset contained publications in a wide range of languages. As several languages, however, were represented only with relatively low numbers of articles or papers, and in order not to misrepresent the research communities these publications stem from, we decided not to take the materials in several languages into account: Polish, Norwegian, Portuguese, Swedish, Danish. As a result, publications written in English, French, German, Italian and Spanish are included in the corpus that is used for the study.

In the next step, a metadata table was created. This table initially included the following data for each publication: id, author, source, type, language, title, year, keywords. Given the considerable number of publications contained in the corpus (5713 texts in total) and the fact that DOIs are not ubiquitous yet for this type of publications, it was necessary to define a unique identifier for each publication that would also serve as the document's filename. This unique identifier consists of the abbreviated name of the conference or journal, the year and a serial number of the publication in the journal or conference. At the same time, in order to facilitate the identification of the publication in contexts other than our study, we decided to keep the original file names. Therefore, two "id" columns were created: primary_id, which contains the project-internal file identifiers, and secondary_id, which includes the original file name.

The next step was to determine which texts relate to CLS. For our convenience, the articles have been divided by language. The texts were checked by title and keywords and categorized as belonging primarily to CLS specifically, or to Digital Humanities more generally. This categorization step had to be done by qualitative inspection, because no suitable keywords scheme exists and existing keywords are not designed for this kind of sub-disciplinary distinction. As a consequence, using just title words or keywords would not have been sufficient to categorize the texts with any substantial degree of precision.

Even when proceeding manually, we found it often very difficult to make these categorizations. When processing the data, we also encountered a number of issues and challenges. For instance, it was sometimes difficult to make the decision regarding CLS / not CLS purely on the basis of keywords and titles, because some keyword combinations are not unambiguous. The field of CLS primarily deals with "the development, the application, and the critique of computational approaches to Literary Studies", according to the "Mission statement" of the *Journal for Computational Literary Studies* (JCLS 2021). Therefore, we looked for words from text analysis, literature and literary studies, text production, and literary history when choosing appropriate keywords. So, the object of study of a relevant to CLS publication must include literary texts in a broader sense, i.e., in addition to prototypical literary genres such as novel, drama, or poetry, it can also include travel reports or fan fiction. Publications, which are potentially relevant to CLS, should also mention algorithmic, statistical, computational or formalized procedures and techniques used to handle literary and textual data. For example, while the term "Stylometry" is mostly a reasonably good indicator for a publication belonging to CLS, it can also appear together with "music" or with "comics / visual media" and in this case, the publication may well be considered to be outside the scope of CLS. Some articles have the term "biography" in their titles or keywords, but whether or not this is an indicator of a publication related to CLS depends on whether the text and the "biography" mentioned as a keyword relate to history, music or literature. Keywords that turned out to be problematic for the categorization process included "manuscripts" (often also in a context, primarily, of digital scholarly editing), "OCR", "handwriting recognition", and "digital edition" (usually an indicator of digital scholarly editing, but in some cases simply indicative of digital editions being used in the process of building a corpus for literary text analysis within CLS). In case of doubt, we followed the strategy to include, rather than exclude, a publication in the CLS category.

In order to make this process and its result transparent, we created the "CLS" column in the metadata table. There, we marked the publications relevant to CLS specifically as "1" and publications relevant to DH more generally as "0". As a result, we obtained a list of texts that we considered related to CLS. This list forms the basis for the smaller subcorpus of CLS publications, made up of 1362 texts, on which further analysis of the texts was carried out. In order to illustrate the structure of our data set the following two graphs were provided.

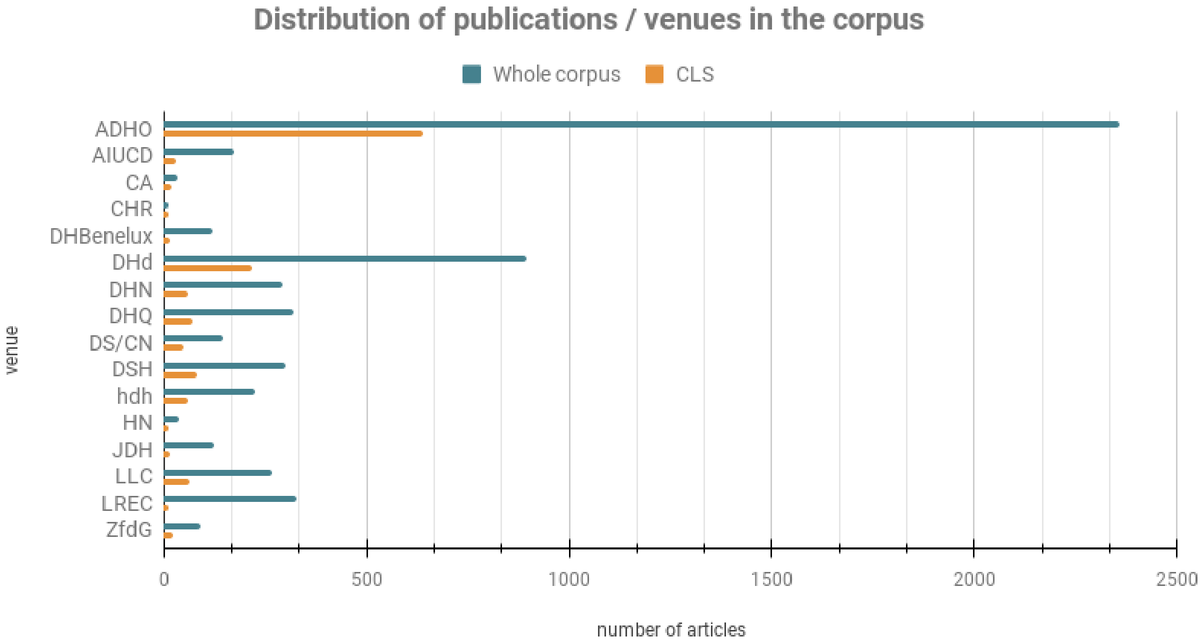

Figure 2. Visualization of publications according to venues and its distribution in the initial corpus and in the CLS corpus.

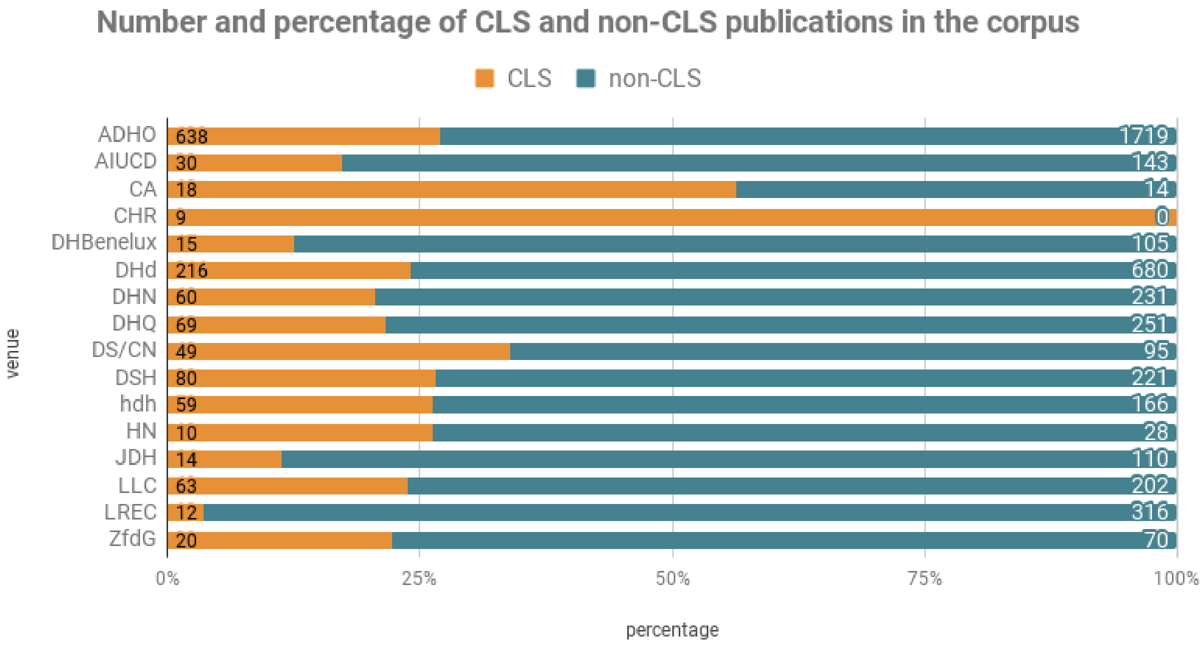

Figure 3. Distribution of CLS and non-CLS publications in the corpus stated as a percentage and absolute numbers. Absolute number of publications are shown on stacked bars.

As it is shown on figure 2, the vast majority of all publications in both whole corpus and CLS subcorpus are taken from the DH Conference organized by ADHO. A large portion of the publications were selected from DHd. About one third of the publications from the entire corpus relates to CLS. This trend is well visualized on figure 3. LREC provided the least CLS-related articles, while all publications from CHR have been used for the CLS-corpus, because in this case, we selected only the articles relevant to CLS from the beginning.

3. Methods

In the following section, we first describe the strategy we followed to define the list of terms designating tools, data formats and methods that we would then search in the corpus. Then we will describe the process we followed to obtain the tables of mentions of tools, data formats and methods in the corpus.

3.1 Delimitation of searched terms

The general challenge is similar for tools, data formats and methods, when it comes to defining the list of terms to take into account: in all three cases, we were concerned with issues of definition and scope (e.g. ‘What is a tool, and what is not?’), granularity (e.g. ‘Do we take the terms for different types of machine learning classifiers into account or not?’), and hierarchical relations between terms (e.g. ‘Do we subsume TEI and RDF under XML or not?’). However, these challenges also have specific nuances in each case, so that we discuss them here one by one.

Tools. There are a large number of research tools and software used in Digital Humanities. Most commonly all programs are called tools, but in our research there is a problem of determining what to call a tool. In studies similar to our own (see those mentioned in the introduction), Python, JavaScript, html, Google, Google books, Twitter, etc. are considered to be tools. Thus, programming languages, markup languages, Internet resources, social networks, etc. are ranked as tools. We find it controversial to classify such resources as tools, since by a tool relevant in our context, we understand programs, packages and libraries that are used for specific tasks in the field of Digital Humanities. We decided, however, to include programming languages in our tool list in order to understand what programming skills are most required in the Digital Humanities. Also, pragmatically speaking, ignoring them later is easier than subsequently adding information on them.

As a starting point, we used the list of tools from the Text Analysis Portal for Research (TAPoR). [5] This is a directory of tools that are used mostly for text analysis by digital humanists. We have edited it according to analyzed and summarized information about the meaning and the purpose of the tool in the context of this study. As a consequence, social media networks such as Twitter, internet browsers and relatively generic project management tools were removed from the list.

Data formats. Digital data carries information and depends by nature on a software that processes it.[6] There is a common understanding of a data format: A data format is a way to store and use (normally within different softwares) digital information (Morrow & Casucci 2019).

[5] See: https://tapor.ca/home.

[6] See: https://guides.library.ucla.edu/c.php?g=180580&p=1186565.

This definition is consistent with our understanding of data format. In order to generate a suitable list of candidate data formats, we consulted, merged and consolidated lists from the following resources: the UCLA library's guide on "Formatting your data", IANUS' recommendations on Data formats, and DARIAH-DE's recommendations.[7] The focus was on the formats used in the Digital Humanities. In addition to a range of file extensions (like "TXT" and "CSV") and text formats, the original list of data formats also includes audio formats and formats related to the music processing, since audiobooks are quite widespread nowadays. Also, we did not exclude database formats from the list, since CLS involves the use of corpus linguistics tools, where textual databases are fundamental. The list also includes image formats as they are often used in Optical Character Recognition, and markup languages such as XML, HTML and XHTML.

Methods. First, we compiled a list of methods. The word "method" has a broad interpretation and is a subject of scientific debate (Johnson et al. 2007). In the context of scientific research, there are three types of approaches to studying a topic, namely quantitative, qualitative, and mixed methods. These approaches are used for collecting, analyzing, and interpreting data in order to understand a subject or phenomenon (Williams 2007). In the TaDiRAH taxonomy, research methods correspond to research techniques and involve foremost capturing, enriching, and analyzing of data, which matches our understanding of methods in the context of CLS.[8] Thus, by methods, we mean the specific set of practices and technical means to achieve the goals of CLS research. The list consists mostly of methods that are used in Computational Linguistics and Digital Humanities, since these are the areas that correlate most with CLS, namely, the tools and scientific methods from these areas are used by researchers in CLS. Our list includes such methods as natural language processing, topic modeling, data mining, text mining, etc. Because our corpus of research literature contains articles not only in English but also in French, German, Spanish and Italian, we needed translations of methods into these languages. The list of CLS methods ultimately used was therefore multilingual. Further, we adjusted the script in accordance with the new task – finding CLS research methods in the corpus. It was based on the method we used for tools and data formats extraction, namely keyword search, where the translation of a method into other languages was added to its English equivalent.

[7] See: https://guides.library.ucla.edu/c.php?g=180580&p=1186565 (UCLA), https://ianus-fdz.de/it-empfehlungen/dateiformate (IANUS) and https://wiki.de.dariah.eu/pages/viewpage.action?pageId=159220082 (DARIAH_DE).

[8] See: http://tadirah.dariah.eu/vocab/index.php?tema=92.

3.2 Ways of counting mentions of terms

In order to identify tools, methods, annotation layers, data and metadata formats and standards used most prominently in the CLS community, we applied basic computational techniques implemented in the form of Python scripts. At the first step we counted occurrences (absolute term frequencies) of the terms of interest from predefined lists (see section 3.1) in our corpus of research articles and abstracts. In order to do this, we considered possible differences in spelling (lower case, abbreviation, variant spellings etc.) of many terms. One example among many is the many ways in which “JavaScript” could be spelled, whether “Javascript”, “Java Script” or “Java script”. We searched separately for each possible spelling variation and then summed up their frequencies. We performed this procedure for each list separately.

To capture the number of documents in which each of the searched terms occurs at least once we binarized the absolute frequencies obtained in the previous step, in effect obtaining an absolute document frequency. The distinction between term frequencies and document frequencies is relevant as it helps to distinguish between terms that are mentioned many times in a small number of different publications and those that are mentioned across a wide range of different publications, sometimes without reaching the same level of total number of mentions of other terms.

In a second step, we divided our corpus in two parts according to the year of publication of articles: an earlier period lasting from 2011 to 2015 and containing 540 publications, and a later period lasting from 2016 to 2021 and containing 822 publications. Then we counted term frequencies and document frequencies for each term for each period. Because the number of articles for each period differs considerably, we transformed absolute frequencies to relative frequencies, both for the term frequencies and the document frequencies. In the case of term frequencies, we used the number of tokens in each subcorpus to obtain the relative terms frequencies. In the case of document frequencies, we used the number of documents in each corpus part to obtain the relative document frequencies.

4. Results

The results of the study are organized as follows. The following subsections each correspond to one of our three focus areas, namely tools, data formats, and methods. The results of each analysis are provided in several steps.

- First, the absolute frequency of mentions of items in the corpus is provided in the form of a bar chart showing the top 30 results (term frequency, including multiple mentions in one single publication) sorted by decreasing frequency.
- Second, the number of articles that contain at least one mention of an item is provided, again in the form of a bar chart showing the top 30 results (document frequency, a very simple indicator of frequency and dispersion).
- Third, we provide the term frequency not for the entire corpus, but split into an earlier and a later period, based on the publication date of the articles: from 2011 to 2015 and from 2016 to 2021.
- Fourth, and finally, we provide the document frequency also split into two periods, similarly using a relative document frequency.

Based on these four perspectives on our data, we are able to identify the most relevant tools, methods, and formats in the selected research articles as well as any large-scale trends of tools, data formats and methods that fall or rise in relevance during the 10-year period we have observed.

4.1. CLS Tools

As the results of the tools analysis show (figure 4), the tools mentioned most frequently in CLS papers are LIWC (Linguistic Inquiry and Word Count), Stylo, Python, Voyant and Mallet. Concerning Stylo, Python, Mallet and Voyant, these results were expected. Python is a very popular programming language in the DH community and in CLS in particular as it is interpreted and offers a lot of useful and easy to implement tools and packages for text processing and text mining. Mallet is the number one tool when it comes to topic modeling, Stylo is a tool that is prominently used in authorship attribution. The Voyant tools is a web-based open-source family of tools that can be used for reading and analyzing texts. These tools offer a lot of instruments for text exploration, such as frequency computing, collocations and context analysis, and many visualization possibilities.[9] Concerning the high frequency of LIWC, these results are a bit surprising, as the mentioned tool is not so typical for the CLS community. LIWC is used to

[9] For further information see: https://voyant-tools.org/docs/#!/guide/about.

classify text along psychological categories and therefore is widely applied rather in psychology or social science than in CLS (Pennebaker et al. 2001). Other frequently mentioned tools are Tesserae, CATMA, Omeka and Gephi. Tesserae is used for exploring intertextual parallels, CATMA is a well known annotation tool and Gephi is a leading visualization and data exploration tool based on Java.[10] Among the mentioned tools, Omeka, which is a platform for digital exhibitions, is rather unexpected.

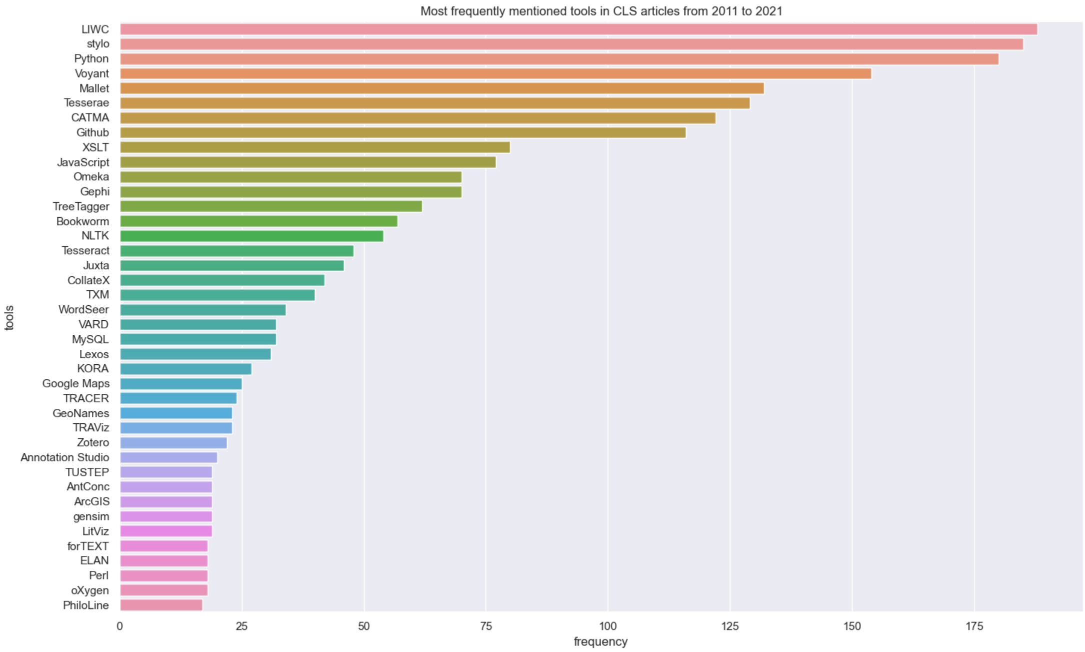

Figure 4. Most frequently mentioned tools in CLS articles from 2011 to 2021. The top 40 tools are shown. The graph shows absolute term frequency.

Comparing these results to the list of tools that are mentioned in the highest number of articles (figure 5), revealed that Python, Stylo and Mallet are still among the top 5. Github and JavaScript (rank 3 and 4 respectively) appeared for the first time among the top 5. The importance of Github can be expected, as this is the largest web service for hosting IT-projects and their collaborative development.[11] The high rank of JavaScript is also not very surprising, as this is another important programming language. While it is primarily designed for web

[10] For further information see: https://gephi.org/
[11] See: https://github.com/about.

development rather than for text analysis, it appears to be used by CLS researchers to create (interactive) visualizations and to present results of analyses on websites.

Another observation is that Voyant tools appear a little bit further below on the graphic. CATMA and Omeka appear in a relatively small number of articles, while LIWC falls down almost to the bottom of the top 30 tools. This observation allows us to conclude that the high position of CATMA, Omeka and LIWC at figure 1 was rather ocasional and can be explained by the high frequency of mentions of these tools in a very small number of articles. Among other tools with the highest document frequency are XSLT, Gephi, TreeTagger and NLTK. We understand that XSLT is actually not a tool, but as we decided to add Python and JavaScript to our tool list, we suppose that XSLT should be included in this list as well. XSLT is a language for transforming XML-documents. TreeTagger and NLTK are widely used tools for language processing. So these results could also be expected.

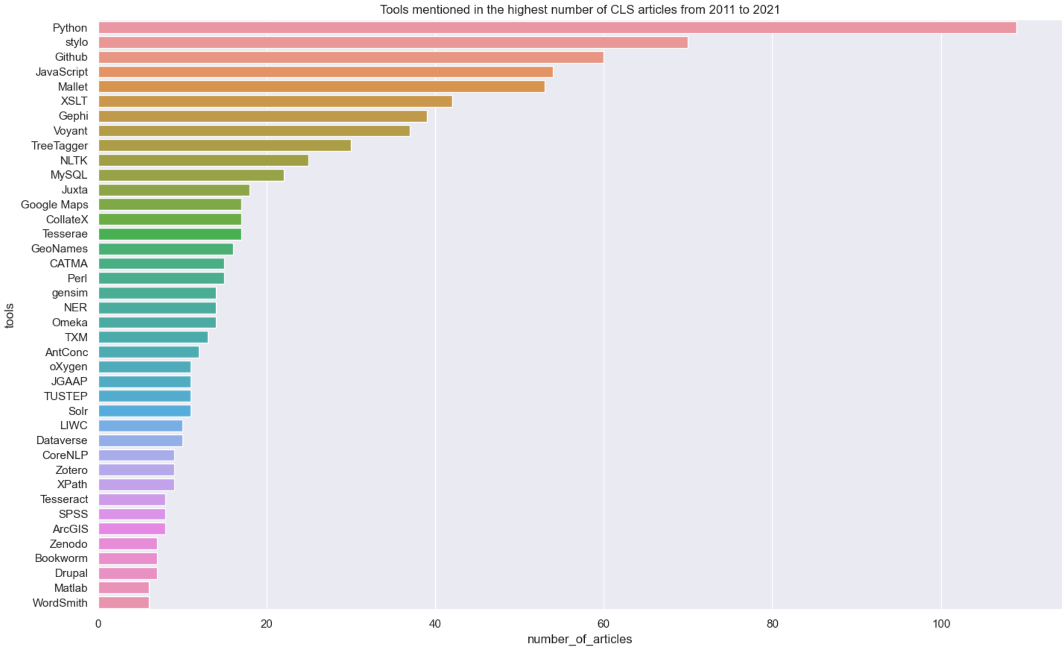

Figure 5. Tools mentioned in the highest number of CLS articles from 2011 to 2021. The top 40 tools are shown. The graph shows absolute frequency.

At figure 6 we can observe the frequency of tool mentions according to two periods: 2011-2015 and 2016- 2021. The tools are sorted according to the total number of mentions. Taking a look at figure 6 shows that LIWC, Python, Stylo, Github, CATMA, TreeTagger, Gephi and Omeka became more popular during the second period (2016-2021). The mentions of XSLT, Voyant, Tesserae and Bookworm decreased significantly during the period of 2016 to 2021 compared to the period of 2011 to 2015. Some tools such as Wordseer, KORA and ELAN are not even mentioned in articles of the second period.

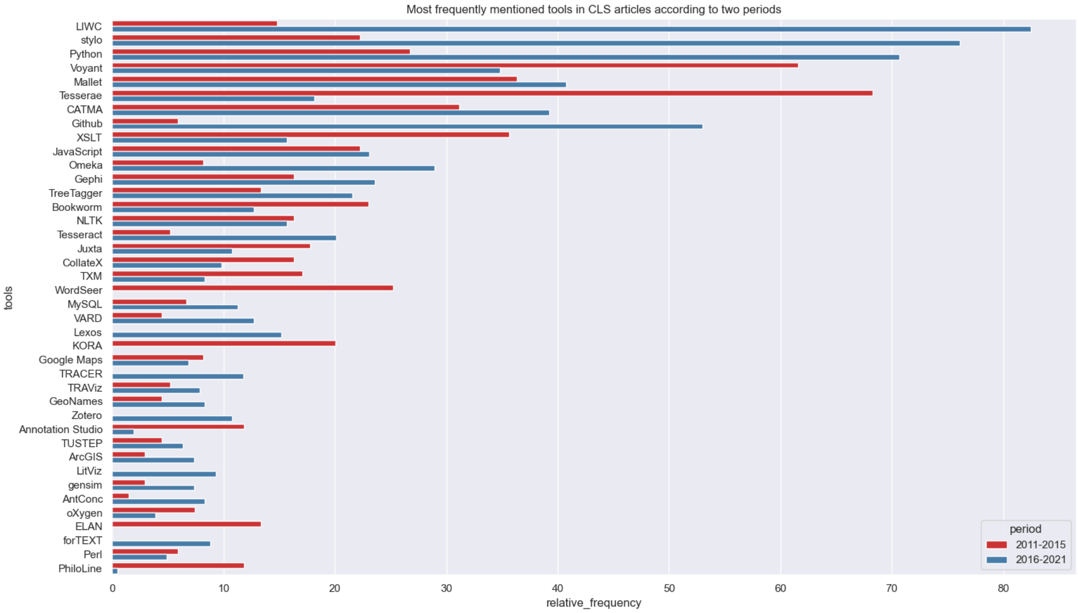

Figure 6. Tools mentioned most frequently in CLS articles according to two periods.The top 40 tools are shown. The graph shows relative frequency (relation to the number of tokens for a certain period). Scale: 1:1 M.

Comparing these results with the document frequency results for the two periods (figure 7), shows that the popularity of Python, Stylo, TreeTagger, Gephi and Github also increased here, during the second period. LIWC and Omeka were also mentioned in a higher number of articles from the second period, but the differences between two periods is not as significant as in the case of term frequency (figure 6). Concerning CATMA, here the document frequency is almost the same for the two periods. When we focus on the tools that lost their popularity during the second period, we notice that results of the previous figure are true only for XSLT and Tesserae. Concerning Bookworm, it appears in almost the same number of articles in the two periods. Interestingly, according to document frequency, Voyant became even more popular during the second period. Some tools like Zotero (first released in 2006), Zenodo (first released under this name in 2015) and Dataverse (launched in 2006) were not mentioned at all during the period of 2011 to 2015.

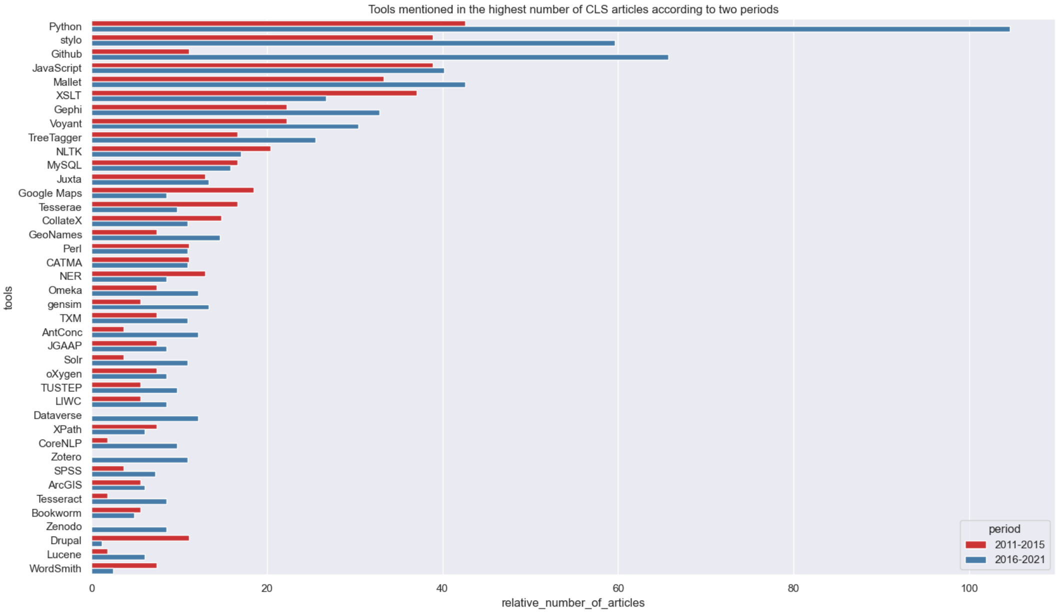

Figure 7. Tools mentioned in the highest number of CLS articles according to two periods .The top 40 tools are shown. The graph shows relative frequency (relation to the number of articles for a certain period). Scale: 1:1 K.

4.2. CLS Data Formats

This section focuses on the various data formats used in CLS research.

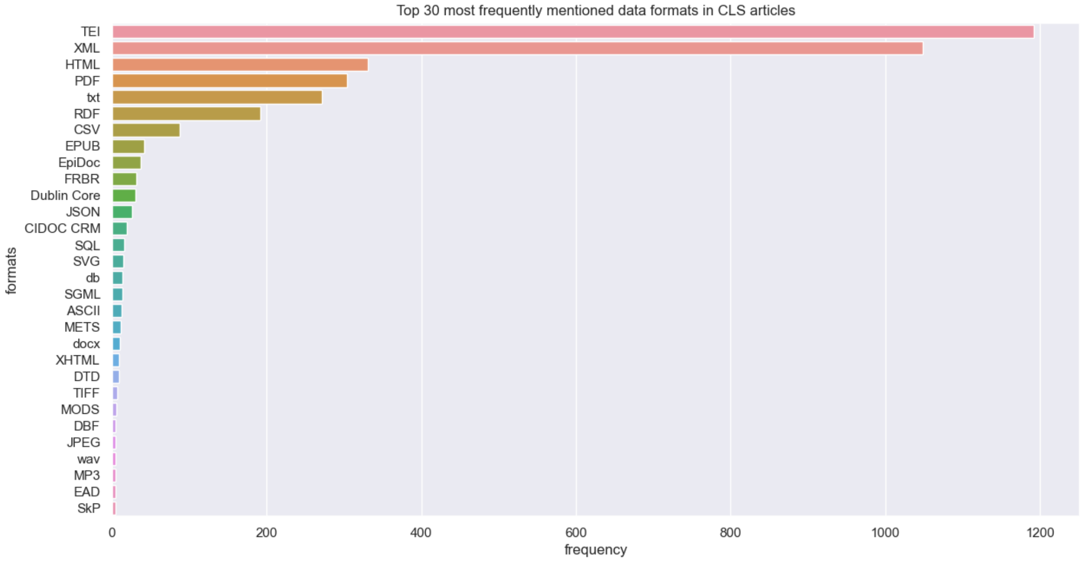

Figure 8. Top 30 most frequently mentioned data formats in CLS publications (absolute term frequency).

Even just a quick glance at the figure above (figure 8) shows that the most frequently mentioned data formats in CLS publications, by a significant margin of their term frequency, are XML and TEI formats (where TEI is in fact an instance of XML-based formats). The same is true, in fact, when we consider the document frequencies (figure 9 below). This is, to some extent, a predictable result, because both TEI specifically and XML more generally are of course standard formats for working with literary textual data (for TEI) or, really, any data (for XML). Also, many of the larger archives and repositories that are frequently-used sources of literary textual data, do provide their texts in XML-TEI.[12] However, the extent to which the XML/TEI format appears to dominate the space is striking. This is especially true given that some of the most popular tools focused on treating textual data (see section 4.1, especially figure 5), either do not rely on or in some cases do not even support textual data provided in an XML and/or TEI-based format: examples include Stylo, MALLET, Voyant, TreeTagger and NLTK. The first tools in the list to rely on data in XML-TEI or to explicitly support XML-TEI are Juxta and

[12] Archives and repositories providing texts in XML-TEI (usually among other formats) include, but are certainly not limited to, the following examples: Théâtre classique (French), Deutsches Textarchiv (German), TextGrid's Digitale Bibliothek (German), Oxford Text Archive (primarily English), Biblioteca italiana (Italian), DraCor (multilingual), ELTeC (multilingual).

CollateX, two collation tools in fact associated primarily with digital scholarly editing rather than CLS more strictly speaking, although there are of course many use cases for collation also in CLS research.

The next group of rather popular formats at ranks 3-5 are, not unexpectedly, HTML, PDF and TXT (plain text). Texts found ‘in the wild’ or in various kinds of platforms, digital libraries or archives will often be available in one of these formats. As most text analysis tools do not specifically support texts provided in PDF or HTML (with the exception of Voyant), we can assume that these formats are mentioned as source formats in CLS research. This hypothesis is confirmed by a look at the occurrences in the corpus. (EPUB, the ebook format that often serves a similar purpose of being a source format in corpus building, can be found a few ranks below in the bar chart.)

Further in the list regarding ranks 6-12, the list includes formats for storing various kinds of data or metadata: RDF, CSV, EPUB (already mentioned), EpiDoc, FRBR, DublinCore, and JSON.[13] It is interesting to note that RDF, the XML-based format associated with Linked Open Data, is rather highly ranked in the list. CSV is a truly multi-purpose format used either for metadata or for linguistically-annotated corpora in one of many so-called ‘vertical’ or tabular text formats. EpiDoc is a markup format specifically designed for epigraphic documents and in fact a fully compatible subset of XML-TEI. FRBR is not strictly speaking a data format, but rather a metadata model used in particular in the area of library and information science. DublinCore is another metadata model used for describing publications and usually implemented in XML. JSON, finally, is not XML-based and another general-purpose data format used to hold key-value pairs of arbitrary content. The two distinct data formats strictly speaking mentioned here, CSV and JSON, are no doubt frequently used because they are entirely content-agnostic.

Finally, one can also observe that there are a number of formats that made it into the frequency range of ranks beyond rank 12, among them formats that support working with media other than text, in particular image data (SVG, TIFF, JPEG, PNG) or music data (MP3, WAV), and formats for working with databases (SQL, db).[14] Finally, further markup formats (SGML, XHTML, EAD), formats for metadata or bibliographic data models (METS) or even ontologies (CIDOC-CRM) can be found here.[15]

[13] CSV means ‘comma-separated values’. EpiDoc stands for ‘Epigraphic Documents’. FRBR stands for ‘Functional Requirements for Bibliographic Records’. DublinCore refers to the ‘Dublin Core Metadata Element Set’. JSON stands for ‘JavaScript Object Notation’.

[14] SQL stands for ‘Standard Query Language’ and ‘db’ simply is the extension conventionally used for database files.

[15] CIDOC-CRM stands for the Conceptual Reference Model’ of the Comité International pour la Documentation’.

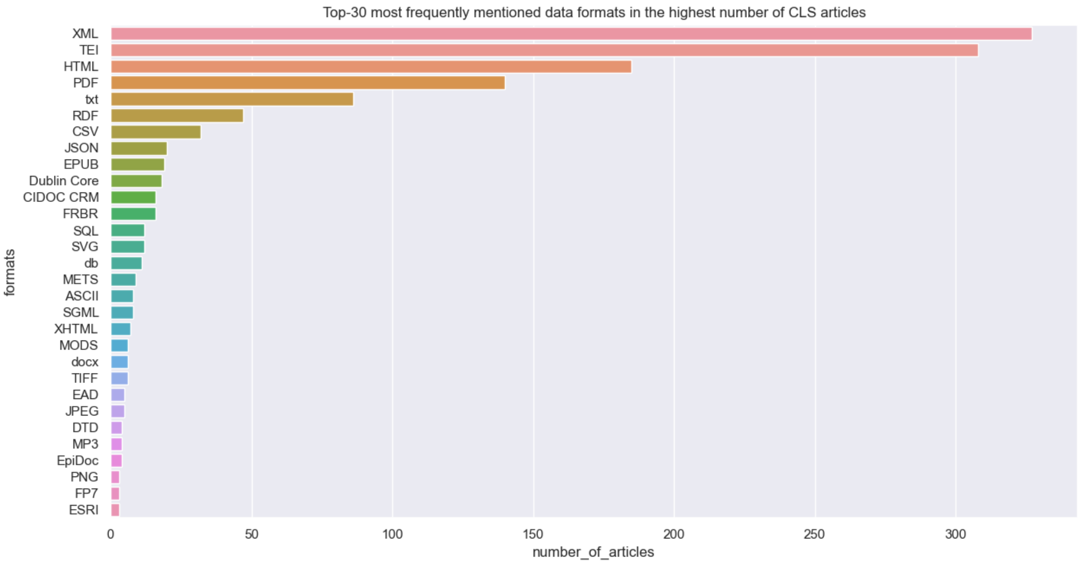

Figure 9. Top 30 most frequently used data formats by absolute document frequency.

In the case of the data formats, there is little difference between term frequencies and document frequencies (figure 9), apart from the fact that the markup format EpiDoc drops considerably down the ranked list (a sign that few articles contain the term but if they contain it, they discuss it extensively) and is replaced by CIDOC-CRM that just about makes it into the top-12 rank. In general, the ranking distribution of formats varies only to a limited extent.

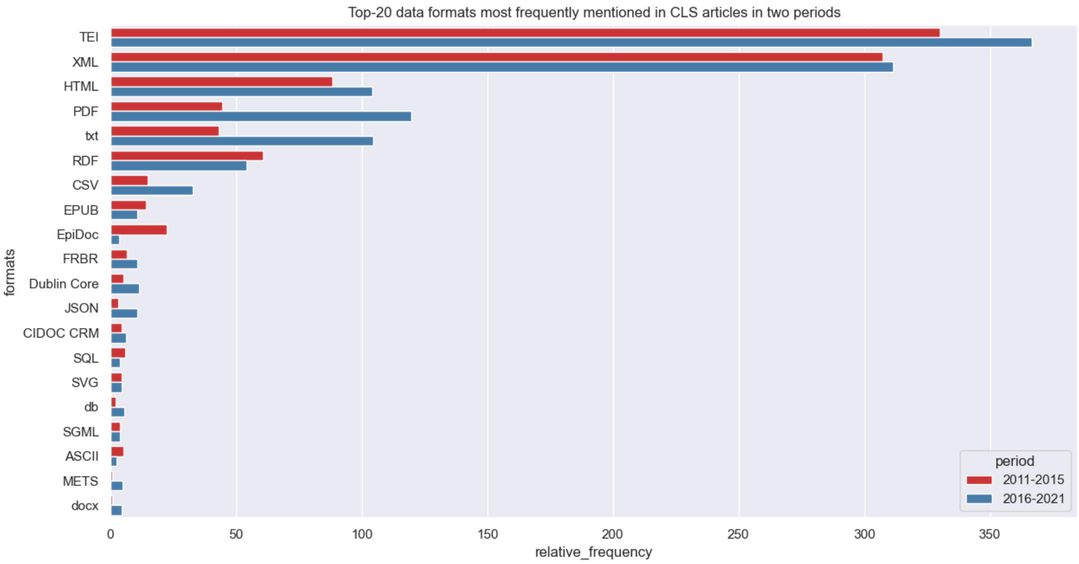

Figure 10. Most frequent data formats in the periods 2011-2015 and 2016-2021, by relative term frequency (frequency per 1 million words).

With respect to the data split into two consecutive time periods, it is interesting to note that the five top-ranked data formats (TEI, XML, HTML, PDF and TXT) all have a higher relative term frequency in the more recent subset of the publications than in the earlier subset. The increase is particularly strong, in relative terms, for PDF and TXT. We do not have a good explanation for this observation. The increase for CSV is similarly dramatic, but from a lower overall level. To our surprise, there is a slight drop in frequency for RDF despite the fact that in our perception, Linked Open Data is playing an increasingly important role in CLS in recent years.

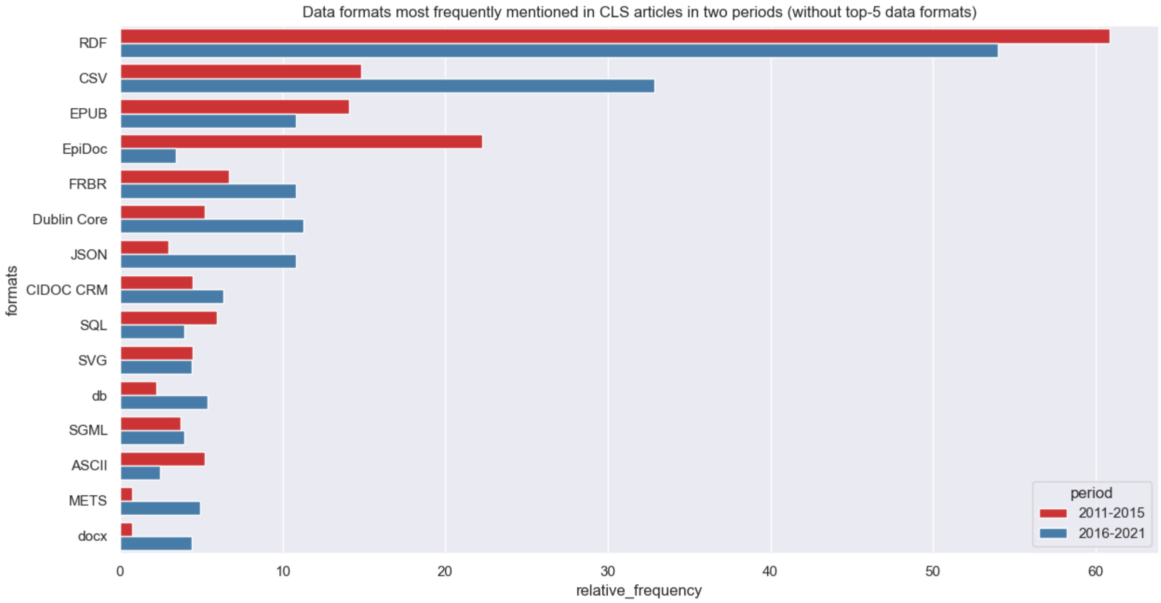

Figure 11. Frequently mentioned data formats in two periods (ranks 6 to 20), according to relative term frequency.

In addition to figure 11, we consider the data again split into two periods, but with a focus on the items of ranks 6 to 20 by relative term frequency. In this view of the group of somewhat less frequent formats overall, we see the very substantial increase of mentions for the CSV format from the earlier to the more recent period. In addition, it is striking that FRBR, Dublin Core and JSON see very clear increases in frequency, albeit at a relatively low level, whereas EpiDoc seems to lose ground in a quite dramatic manner. This may be due to disciplinary shifts, of course, or even be an artifact of the corpus composition, rather than signal an actual decrease in importance of EpiDoc in Digital Humanities overall.

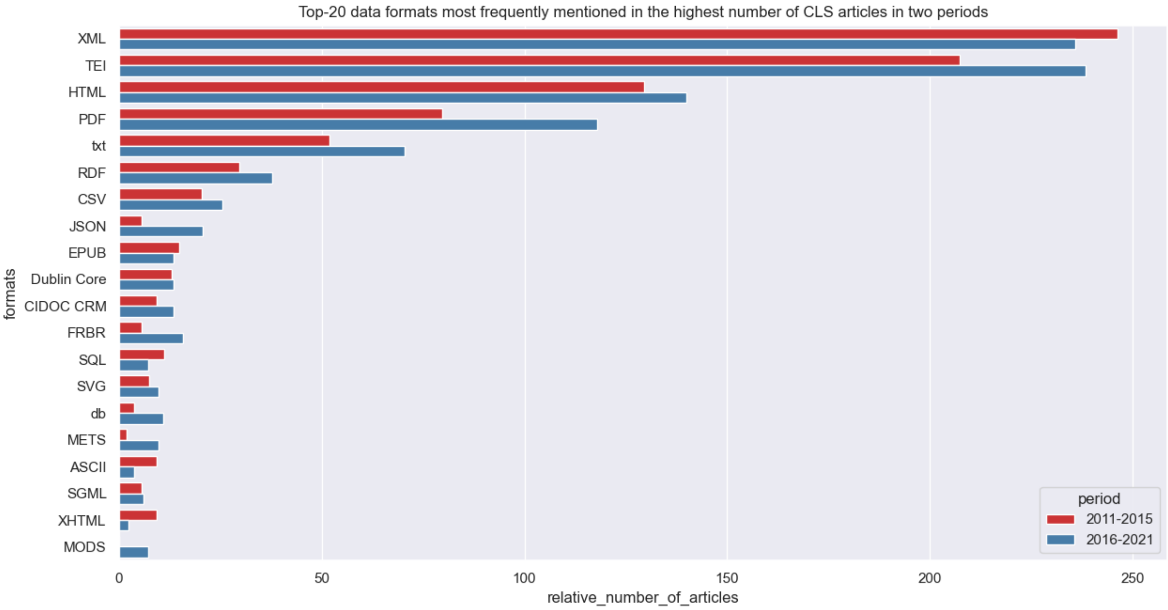

Figure 12. The top-20 most frequent data formats, split by two time periods (2010-2015 vs. 2016-2021), based on relative document frequency.

Compared to the relative term frequencies, the picture seen here (figure 12) for relative document frequencies is only slightly different. The frequency of mentions of XML decreases, while the frequency of mentions of TEI increases. The remaining five top-ranked formats show a trend of increasing importance that is, however, less pronounced especially for PDF and TXT when compared to the relative term frequencies (figure 10).

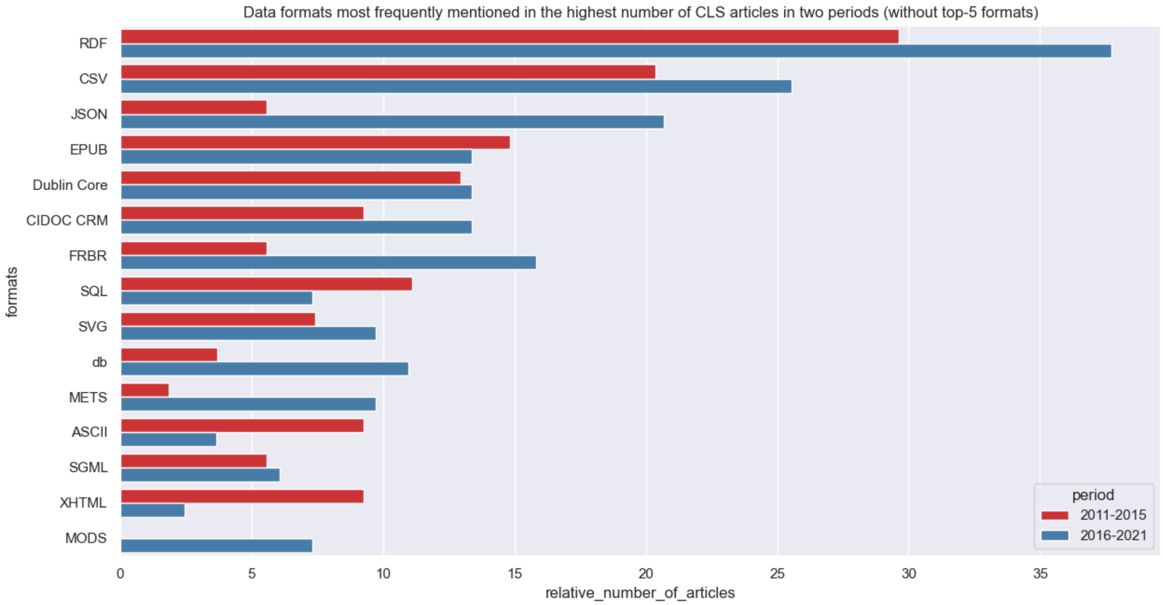

Figure 13. Most frequently mentioned data formats (ranks 6 to 20) for the two time periods (2010-2015 vs. 2016-2021), based on the relative document frequency.

If we look at the results for the ranks 6 to 20 in our graph, we can see again that the strongest increase is present for JSON (whose frequency basically quadruples), but other formats also increase in relative document frequency, like RDF, CSV, CIDOC-CRM and FRBR, db and METS. Conversely, SQL, ASCII, and XHTML are found in a larger proportion of articles in the earlier period compared to the later period.

4.3. CLS Methods

The next step in our research was to identify the methods that are most commonly used in Computational literary studies. The first chart (figure 11) demonstrates the absolute frequency of CLS methods references in the corpus.

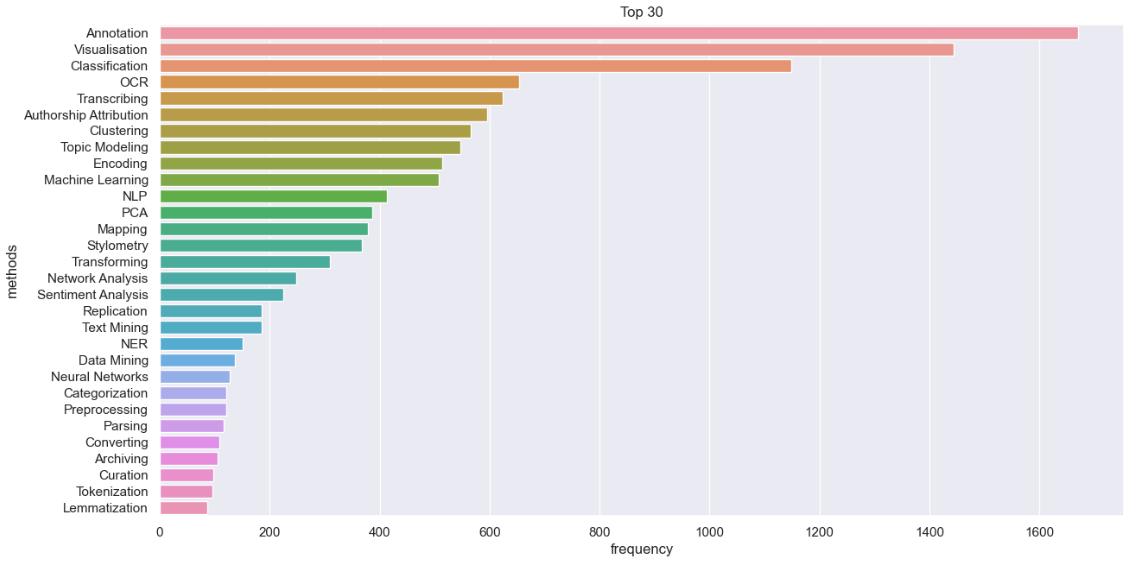

Figure 14. Top 30 of the most frequent methods in the CLS corpus.

The graph (figure 14) shows that the leading methods are annotation, visualization and classification, which are quite “classical” methods for analyzing literary texts. These results correspond with the results of the data format frequency analysis shown in graphs on figures 8 and 9, where the leading positions are taken by markup languages such as XML, TEI and HTML. In terms of annotation these data formats play an important role and are widely used as a type of data. CATMA, an annotation tool, is also among the top tools mentioned in the corpus, which proves the importance of annotation as a CLS method. The importance of visualization is supported by for example Voyant and Gephi, which support text and graph data visualizations, respectively. Specific CLS techniques that are on the list of method extraction results include OCR, authorship attribution, topic modeling, principal component analysis (PCA) and stylometry. Machine learning and NLP techniques, such as text mining, neural networks, NER, data mining, sentiment analysis, are also often mentioned in the corpus. Interesting to note that Stylo, a tool which deals with stylometry, is on the top of the tool list according to term frequency, while this method showed an average result. However, the method of authorship attribution, which is in principle similar to stylometry, is one of the leading techniques according to our results. Tokenization and lemmatization are inferior in frequency to other methods of text analysis.

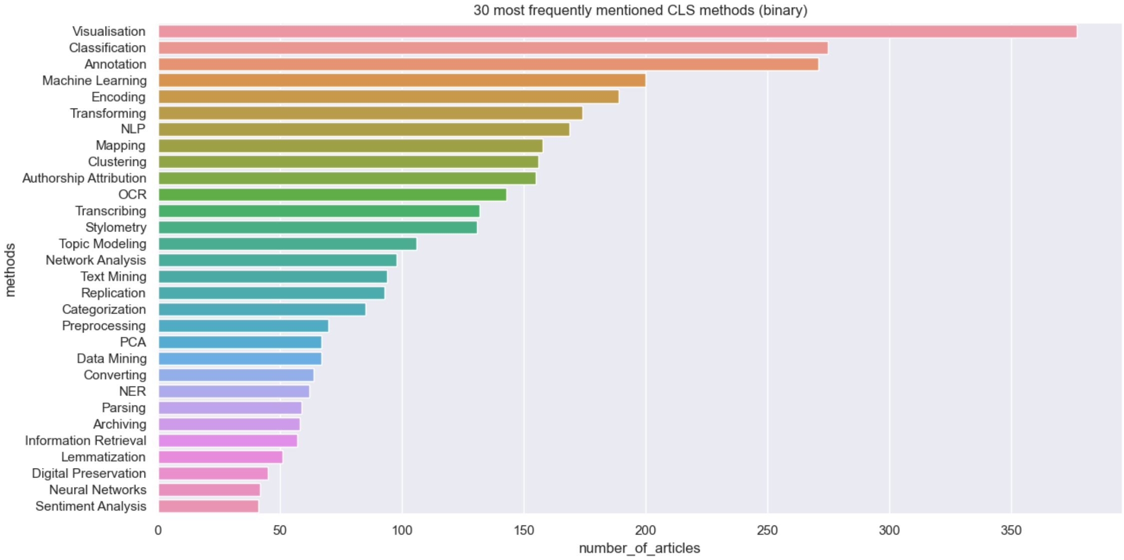

Figure 15. The most frequently mentioned methods in the highest number of articles.

If we look at the list of methods which are frequently mentioned in the highest number of CLS articles, it must be stated that the results are slightly different (figure 15). The method of visualization comes first, overtaking annotation and classification. This method is mentioned in the absolute majority of articles (more than 350). Machine learning and its methods (NLP, text mining) prove to be significant in the CLS area, as they are mentioned in a large number of articles from the corpus. Such methods as OCR, authorship attribution, stylometry, topic modeling have also proved to be important methods in a large number of publications. Surprisingly, sentiment analysis and neural networks are among the last on the list, whereas in the absolute frequency graph these methods play a more important role.

A comparison of the two periods (figure 16) showed the following results. Annotation, transcribing, authorship attribution, PCA were largely more frequently mentioned in the early 2010s than closer to 2020s. We could notice a slight difference in occurrences of the first three leading methods. The absolute frequency is more or less evenly distributed throughout the corpus, regardless of the year of the studies.

The second time period is characterized by more frequent references to methods such as OCR, clustering, topic modeling, NLP, sentiment analysis, network analysis, transforming and NER.

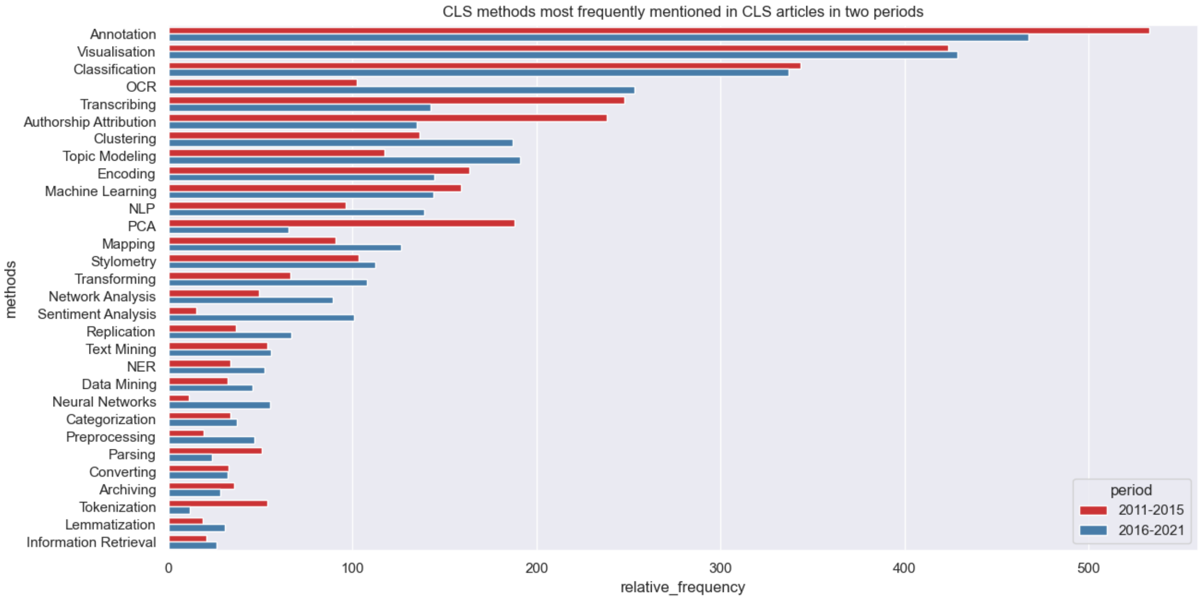

Figure 16. Frequency of CLS methods in two periods, top 30 are shown.

The graph which shows the frequency of CLS methods in two compared time frames (figure 16) demonstrates that visualization, annotation and classification were used more often in early articles, and less frequently over time. Such methods as authorship attribution, transcribing, encoding, mapping, clustering are mentioned much more frequently in earlier articles.

It is noticeable that there is a tendency for almost all methods to be mentioned earlier more often than in more recent articles. Nevertheless, more modern methods that belong to machine learning can be seen more often in recent articles. This proves that the CLS field is increasingly using more modern ways of analyzing text, not limited to classical tools for literary analysis. NLP, OCR, topic modeling, network analysis, replication, preprocessing have been mentioned more frequently in the newest publications, which indicates an increase in the importance of these methods over time.

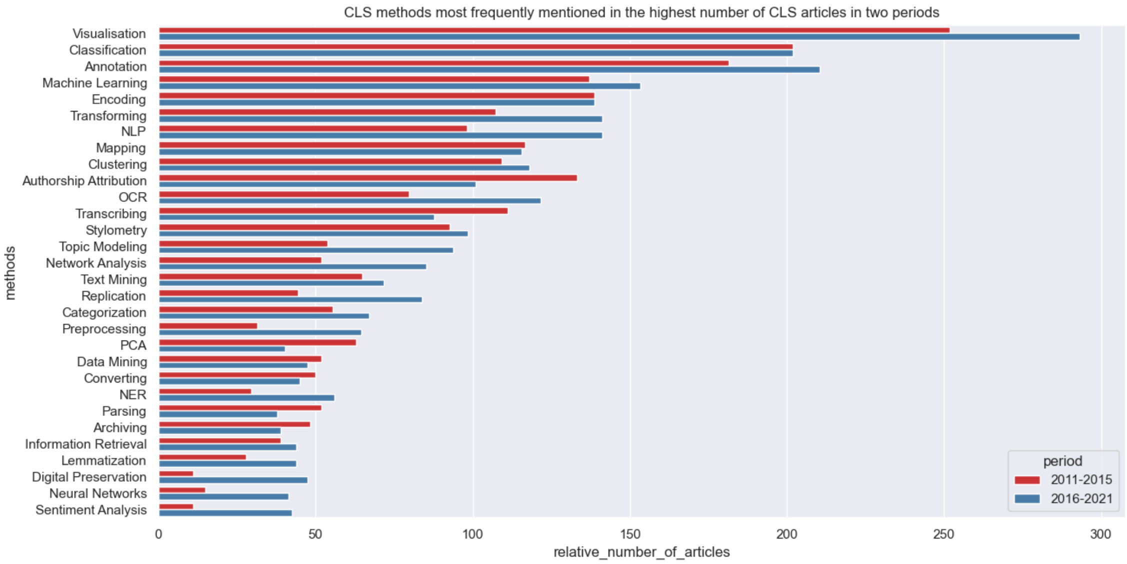

Figure 17. Relative frequency of CLS methods occurrences for two periods, top 30 are shown.

Interestingly, all of the top methods become relatively less important. As these are relative document frequencies, this “loss” must be made up by a large number of less important methods that become mentioned more often. This might be a signal for the increasing differentiation and specialization within the CLS.

4.4 Observations across domains

An interesting question is whether our findings for tools, data formats and methods appear to be in alignment with each other or not. In other words: Are the most frequently mentioned tools also supporting the most frequently mentioned methods? Are the most frequently mentioned data formats supported by the most frequently mentioned tools? The picture in this regard is not entirely clear, especially because some functionality might be covered by general-purpose programming languages like Python and JavaScript and their many libraries and packages. But some observations can be made.

First of all, the dominance of XML-TEI for textual data, and of other XML-based formats for other data (such as RDF) does not appear to be fully reflected in the tool landscape. The most frequently mentioned tools can for the most part work with texts encoded in XML-TEI, but do certainly not require this kind of input data (Stylo is a point in case). Other tools don’t handle XML-TEI input very gracefully at all (MALLET, for example). An exception appears to be Voyant, which is exceptionally flexible in the range of data formats it accepts. With regard to other frequently mentioned tools, like NLTK and TreeTagger, they require input in plain text format. It is important to mention that many of the Python and JavaScript packages and libraries require plain text as input as well. According to the term and document frequency in our corpus, plain text (TXT) appears only at 5th position, after PDF and HTML, which actually are source and not input formats. So we can conclude that crucial tools used in the CLS community do not accept input in a very important for the community format, namely XML-TEI.

When we compare the lists of the most frequently mentioned tools with the most frequently mentioned methods, we observe more conformity. The dominance of such methods as annotation, visualization, classification, machine learning, NLP, authorship attribution, clustering, stylometry or topic modeling corresponds with the dominance of following tools: Python (visualization, classification, annotation), Stylo (clustering, authorship attribution, stylometry), Mallet (topic modeling), Gephi (visualization), TreeTagger (annotation, NLP), NLTK (NLP), CATMA (annotation).[16]

In fact, we believe that working towards a stronger and more programmatic alignment between tools, data formats and methods might not only be a good strategic agenda point for CLS INFRA, but also a sign of a growing maturity of the field of Computational Literary Studies more generally.

[16] As we have some very general methods like classification, annotation and also specific methods like OCR or topic modeling, we did not take in this comparison the exact rank of each method and tool into account.

5. Conclusions

To conclude our report, we would like to first summarize our key findings with a focus on the conclusions we can draw from our more detailed results for the development of training opportunities within CLS INFRA (as developed notably in WP4) and for the requirements analysis that will inform infrastructure development in CLS INFRA (in particular in WP6, 7 and 8). Beyond CLS INFRA, these conclusions can certainly also be applied to the CLS community more generally, for example when a training programme decides on particular workshop topics or when an individual researcher decides on which of the many possible tools for text analysis in CLS to learn next. Beyond this, we also reflect on some of the limitations of this study and propose ideas for future work that, at least in part, may serve to alleviate some of the limitations of the present study.

With regard to tools, our analysis shows that Python is the most important programming language for the CLS community. Github is a central platform when it comes to data storage, version control and collaborative code writing. Stylo and Mallet are crucial tools when it comes to stylometry or topic modeling respectively. Other important tools concern language processing or annotation like TreeTagger or NLTK. CATMA did not lose its popularity during the last years and Gephi became an even more important visualization tool in the CLS community. Concerning infrastructure requirements, we have some very specific tasks in the community like topic modeling, literary annotation or authorship attribution and a rather limited number of tools, that can solve these tasks. With regard to training, we consider a high importance of focusing on Python (and key libraries like pandas, scikit-learn, spaCy and/or Stanza), Github / Gitlab and different annotation tools. Also we see a need of being familiar with more specific tools that deal with crucial tasks of the CLS like topic modeling, authorship attribution, word embeddings and stylometry.

With respect to the data formats, our analysis clearly shows that for the CLS community, XML-TEI is almost as fundamental as it is in the community of digital scholarly editing. In addition, key data formats CLS researchers should be familiar with are CSV and JSON, whether for annotated textual data or for metadata. Finally, although they are not strictly speaking separate data formats, one can note the importance of various metadata standards and markup languages that are often but not necessarily expressed in XML. CLS researchers should also be familiar with those. In terms of infrastructure requirements, we believe this means that an infrastructure for CLS should be able to handle texts provided in XML-TEI. In terms of training, it means that a focus on XML-TEI, CSV and JSON as well as on conversion routines from various markup formats from and to XML-TEI can be said to be of considerable importance.

Regarding methods, CLS publications pay special attention to annotation, visualization, and classification. It should also be noted that such a research technique as machine learning is also important for literary research, as its algorithms are widely used for text analysis, including in the CLS field. In addition, modern literary studies include the use of progressive techniques such as NLP, NER, topic modeling, network analysis, neural networks, and sentiment analysis, which show a trend toward wider application. This tendency shows an expansion of the range of quantitative literary research methods and the emergence of more modern computational techniques. This should be taken into account in the structure of CLS training and for the CLS infrastructure as well.

We believe to have provided a solid empirical basis to our conclusions, as they are based on an analysis of more than 1600 publications in the CLS field. However, there are of course a number of limitations to our study that should be taken into account when making decisions based on our study. For example, many languages are of course missing, not just European ones, but also all languages other than English used e.g. in the Asia-Pacific region of the world. In short, our study does not and cannot pretend to present the complete picture of CLS research. On a more technical level, we have of course to contend that the category assignments are a considerable and mostly pragmatic simplification, beyond probably not being 100% accurate as it stands. We also have not attempted to solve the problem of hierarchical relations between terms, as in the case of, for instance, PCA or tSNE,[17] clustering and machine learning (with the latter terms encompassing the former terms) or EpiDoc, RDF, TEI in relation to XML (where all of the former formats are expressed in XML).

With respect to future work, we would of course have liked to include more texts and, in particular, texts in more languages than we have been able to include at this time. As a task not so much for ourselves, but rather for the Digital Humanities community more generally speaking, we believe that more consistent, more structured and more widespread practices with respect to assigning keywords to publications would be highly desirable if we are to understand the internal structure of the field of Digital Humanities that undoubtedly differentiates itself into multiple subfields as it continues to grow. In addition, and based on such keywords, a sharper and more decisive delimitation between publications assigned to the CLS category and those assigned to the more general DH category would be useful. Or rather, but this is really more a more general methodological point, more fine-grained and gradual category assignments, allowing for multiple allegiances and hybrid documents to be appropriately referenced, and methods allowing to calculate with such categories, would be desirable. This might be

[17] tSNE stands for t-distributed stochastic neighbor embedding and is a dimensionality reduction technique that can be used as an alternative to the more established PCA, which stands for Principal Component Analysis.

something we can’t obtain based on metadata alone, but on an analysis of full texts such as the one brilliantly displayed by Lehmann and Burghardt 2019. Finally, it could be interesting to expand our analysis to the entire corpus and compare results regarding CLS only (as the focus has been in this report) with results concerning Digital Humanities more generally.

References

Barbot, Laure, Frank Fischer, Yoann Moranville, and Ivan Pozdniakov. 2019. “Which DH Tools Are Actually Used in Research?” *Weltliteratur*. https://weltliteratur.net/dh-tools-used-in-research/.

Borek, Luise, Canan Hastik, Vera Khramova, Klaus Illmayer, and Jonathan D. Geiger. 2021. “Information Organization and Access in Digital Humanities: TaDiRAH Revised, Formalized and FAIR.” In *Information between Data and Knowledge. Information Science and its Neighbors from Data Science to Digital Humanities. Proceedings of the 16th International Symposium of Information Science (ISI 2021)*, 74:321–32. Glückstadt: Werner Hülsbusch. https://epub.uni-regensburg.de/44951/.

Borek, Luise, Jody Perkins, Christof Schöch, and Quinn Dombrowski. 2016. “TaDiRAH: A Case Study in Pragmatic Classification.” *Digital Humanities Quarterly* 10 (1). http://www.digitalhumanities.org/dhq/vol/10/1/000235/000235.html.

Borgman, Christine L. 2007. *Scholarship in the Digital Age: Information, Infrastructure, and the Internet*. Cambridge, Mass: MIT Press.

DARIAH Wiki. 2021. *Empfehlungen für Forschungsdaten, Tools und Metadaten in der DARIAH-DE Infrastruktur.* https://wiki.de.dariah.eu/pages/viewpage.action?pageId=159220082.

JCLS. 2021. “Mission statement”. *Journal of Computational Literary Studies*. https://jcls.io/site/mission/.

Johnson, R. Burke, Anthony J. Onwuegbuzie, and Lisa A. Turner. 2007. “Towards a definition of mixed methods research.” *Journal of mixed methods research* 1.2, 112-133.

Kitchin, Rob. 2021. *The Data Revolution: Big Data, Open Data, Data Infrastructures and Their Consequences*. 2nd edition. Thousand Oaks: Sage Publications Ltd.

Luhmann, Jan, and Manuel Burghardt. 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): 148–71. https://doi.org/10.1002/asi.24533.

Morrow, Anne, Casucci, Tallie. 2019. *Preserving and Disseminating Emerging Forms of Digital Scholarship in Academic and Research Libraries: The EDS Report.* doi: 10.26052/0G7R9EPF43.

Moulin, Claudine, Arianna Ciula, and Julianne Nyhan. 2011. Research Infrastructures in the Digital Humanities. Science Policy Briefing 42. Strasbourg: European Science Foundation. https://www.esf.org/fileadmin/user_upload/esf/RI_DigitalHumanities_B42_2011.pdf

Pawlicka-Deger, Urszula. 2021-2022. *DH INFRA. Digital humanities, infrastructure and knowledge production*. https://dhinfra.org/.

Schäfer, Felix F. 2016. *IT-Empfehlungen für den nachhaltigen Umgang mit digitalen Daten in den Altertumswissenschaften.* IANUS. https://ianus-fdz.de/it-empfehlungen/dateiformate.

Svensson, Patrik. 2016. *Big Digital Humanities: Imagining a Meeting Place for the Humanities and the Digital*. Ann Arbor MI: University of Michigan Press. http://dx.doi.org/10.3998/dh.13607060.0001.001.

UCLA Library. 2021. *Data Management for the Humanities. File Formats and Software.* https://guides.library.ucla.edu/c.php?g=180580&p=1186565.

Williams, Carrie. 2007. *"Research methods." Journal of Business & Economics Research* 5.3. https://doi.org/10.19030/jber.v5i3.2532.