future internet

MDPI

Article

Digital Qualitative and Quantitative Analysis of Arabic Textbooks

Francesca Fallucchi [1,2,*], Bouchra Ghattas [1], Riem Spielhaus [1,3] and Ernesto William De Luca [1,2]

[1] Leibniz Institute for Educational Media | Georg Eckert Institute, Freisestraße 1, 38118 Braunschweig, Germany; bouchra_ghattas@hotmail.com (B.G.); spielhaus@gei.de (R.S.); deluca@gei.de (E.W.D.L.)
[2] Dipartimento di Scienze Ingegneristiche, Università degli Studi "Guglielmo Marconi", 00193 Roma, Italy
[3] Institute of Arabic and Islamic Studies, Georg-August-Universität Goettingen, 37077 Goettingen, Germany
* Correspondence: fallucchi@gei.de

Abstract: Digital Humanities (DH) provide a broad spectrum of functionalities and tools that enable the enrichment of both quantitative and qualitative research methods in the humanities. It has been widely recognized that DH can help in curating and analysing large amounts of data. However, digital tools can also support research processes in the humanities that are interested in detailed analyses of *how* empirical sources are patterned. Following a methodological differentiation between close and distant reading with regard to textual analysis, this article describes the Edumeres Toolbox, a digital tool for textbook analysis. The Edumeres Toolbox is an outcome of the continuous interdisciplinary exchange between computer scientists and humanist researchers, whose expertise is crucial to convert information into knowledge by means of (critical) interpretation and contextualization. This paper presents a use case in order to describe the various functionalities of the Edumeres Toolbox and their use for the analysis of a collection of Arabic textbooks. Hereby, it shows how the interaction between humanist researchers and computer scientists in this digital process produces innovative research solutions and how the tool enables users to discover structural and linguistic patterns and develop innovative research questions. Finally, the paper describes challenges recognized by humanist researchers in using digital tools in their work, which still require in-depth research and practical efforts from both parties to improve the tool performance.

Keywords: digital humanities; quantitative and qualitative research; textual analysis; Arabic language

Citation: Fallucchi, F.; Ghattas, B.; Spielhaus, R.; De Luca, E.W. Digital Qualitative and Quantitative Analysis of Arabic Textbooks. *Future Internet* **2022**, *14*, 237. https://doi.org/10.3390/fi14080237

Academic Editor: Massimo Cafaro

Received: 4 July 2022
Accepted: 26 July 2022
Published: 29 July 2022

1. Introduction

In the past few decades, digital tools have become increasingly important for supporting in-depth research. However, in order to unlock the dormant potential of digital technologies to improve interdisciplinary scientific processes, the exact requirements of humanist researchers need to be scrutinized and given precise solutions. One example of such an operating principle has been applied by the Leibniz Institute for Educational Media | Georg Eckert Institute (GEI). The Institute promotes and supports research into textbooks and educational media from a historical and cultural studies perspective. Furthermore, it provides several research infrastructures, such as its renowned research library with a collection of international school textbooks, and its various dedicated and customized digital information services. The GEI also employs and implements research infrastructures to support Digital Humanities (DH) researchers (http://gei-digital.gei.de (accessed on 3 July 2022)). Automatic image and text analysis, linguistic annotation of texts, and data visualization are digital methods that allow textbook researchers to investigate diverse questions and document their research process.

Digital tools and data analysis expand the researcher's options when examining or studying a particular subject. Combining digital methods with other research tools in a specific discipline enables humanists to explore information and answer research questions. The use of digital tools has the potential to transform research, offering novel possibilities

for the creation, analysis, and sharing of knowledge. As a specific genre, textbooks lend themselves to multiple text analysis (deductive and inductive) and comparisons in terms of subjects, historical periods, states, regions, or languages, which employ quantitative and qualitative modes of analysis and description and can therefore be facilitated by digital tools in a profound way. Digital tools generally provide very useful functionalities to annotate (code) different features of the text, for example, using different colours and underlining styles. Furthermore, they allow users to add comments, key words, codes, or memos to relevant text passages, images, or whole documents through sustainable and collaborative annotation.

The increased availability of digitized texts and digital resources accessible through digital technologies brings challenges and opportunities with regards to textual analysis. The literary historian and theorist Franco Moretti approached these by differentiating between close and distant reading. Close reading refers to the detailed analysis of a text or corpus, which can follow a diverse set of methods [1,2]. These analysis techniques include, for example, stylometric analysis, which aims to identify the style of an author or genre by comparing it to others, identifying and documenting the characteristic traits that distinguish it. Distant reading, on the other hand, is an innovative approach to literary analysis introduced by Franco Moretti at the beginning of the twenty-first century in the context of digital transformation [3–5]. He adopts a quantitative approach to literary studies, arguing that thousands of documents may need to be analysed in order to get an accurate understanding of what comprises a literary genre. Following this, a combination of the close and distant reading approaches has been developed [6–11].

The Edumeres Toolbox that this paper describes in the following pages responds to the fundamental needs of today's DH researchers and provides all the functionality needed to manage a digital library through a web application. The system also provides specific functionalities to allow researchers to carry out close reading vs. digital close reading, and distant reading visualizations, which are characterized by quantitative and qualitative methods [12].

The remaining part of the paper is organized as follows: Section 2 describes current research trends in the domain of DH. Section 3 presents the use of digital tools for textual analysis. Section 4 shows how content analysis has been implemented in the Edumeres Toolbox with a case study of a collection of Arabic textbooks, and the use of the content analysis functionalities is explained in Section 5. Section 6 concludes the paper with a review of the remaining challenges and opportunities for DH.

2. Related Work

Quantitative and qualitative methods employed in the DH field ensure a number of opportunities and challenges. Both these research methods can be expanded in scope, depth, and sophistication. The workflow of qualitative and quantitative research steps can be formalized, and reliable digital archives can provide support. Human interpretations can be used for training and learning and even for future predictions and post-correction.

Close reading can be defined as the in-depth interpretation of a text passage, which determines the themes by investigating events, names, and concepts; their development and interaction; the use of terms and phrases; and the text's structure and style. Whereas with close reading one can read the source text without dissolving its structure, distant reading does the exact opposite, moving from observing the textual content to visualizing the global characteristics of one or more texts with the aim of generating an abstract view of the text [13].

According to Moretti's ideas, close reading must be limited to the provision of a detailed explanation of the data obtained after subjecting the texts to distant reading, independent of any subjective interpretation. This forces the researcher to act in a mechanical manner, giving up their intellectual autonomy.

Therefore, most supporters of distant reading criticize Moretti's vision and propose a compromise between close and distant reading.

Maurizio Ascari, for example, writes: "Moretti's quantitative approach is the child of the computer age. Far from questioning the usefulness and potential of the DH, I would like to emphasize the need for a combined approach" [14].

Andrew Piper, moreover, emphasizes that close reading should not act as an opposing tool to distant reading and serve only as a confirmation and explanation of the data obtained: The critical contribution made by the subjective interpretation carried out by humans rather than by machines must still play a decisive role [15].

The terms "quantitative" and "qualitative" are understood, shared, written about, and ultimately used in a much more systematic way in the DH, where they are often taken to express a simple distinction between "something than can be computed" versus "something that cannot".

Direct access to the source texts is important for humanist researchers when they conduct research. Visualization techniques are key when it comes to analysing the data provided [16,17]. For example, Bradley emphasizes the need to have a visualization technique that does not destroy the original text in the process [7]. In [8], the authors discuss how distant reading can direct the reader to a part of document that may deserve further investigation and cannot be used to replace close reading. Beals [6] defines the following question: "In an age where Distant Reading is possible, is Close Reading dead?".

A better balance between the "how" and the "what" of DH for a qualitative shift towards an interest in also making sense of and understanding information is discussed in [18].

Over the past decade, the use of computer-assisted qualitative data analysis software (CAQDAS), software that provides a set of data analysis tools [19], has become commonplace in qualitative research. But how is qualitative research in the DH currently being conducted in practice?

In [20], Bosch describes the three core methodological components that characterize systematic qualitative research: 1. "constant comparison" [21,22], 2. "analytic induction" [23], and 3. "theoretical sensitivity" [24]. A mixed-methods approach is often preferred, in which results from the holistic, systematic, and hermeneutic methods based on selection, comparison, categorization, theory development, and synthesis are included [20,25–27].

The advantages and opportunities that arise from the use of these algorithms are clear. There are two questions when approaching the problem with these digital methods, in accordance with [18]: 1. Do algorithms help in the interpretation of implicit meaning in large amounts of resources? 2. How, using these algorithms, can we best find or develop relevant, plausible, and accurate empirical and theoretical material [28]? The answers to these two questions open up several challenges. Can qualitative methods risk overtaking quantitative ones?

Jockers [29] uses the terms macroanalysis and microanalysis to denote distant and close reading, respectively. He suggests a mixed approach for a better and enhanced understanding of literary documentation, where distant and close reading work in tandem.

Schuster and Dunn wrote *The Routledge International Handbook of Research Methods in Digital Humanities* [9] to describe both quantitative and qualitative research trends in the DH, in which they discuss methods applied and what the future directions of research methods in DH might look like.

Several steps are necessary to enhance the potential of combining close and distant reading research in the DH. We believe that the analysed methods, algorithms, and systems present a potential direction for the combination of the two approaches, which includes explicit interpretations of the content and experimentation with the functionalities for large research questions. The process of interpretation must be simplified, explicit, and open to control. The importance of close reading should also not be neglected. Distant reading is certainly different from reading by a human reader, whose fallibility and vulnerability are crucial to the results of the analysis. Even machines may be fallible due to bugs or processing errors, but these remain purely mechanical failures, whereas individual

human thinking will always lead to a different interpretation with each analysis. However, processing is not reading: It is an automatic, repetitive process that aims for precision. Therefore, close reading should rather be understood as a tool that aids the computational power of distant reading.

Qualitative research in the DH holds great promise if the right steps are taken to respond to the challenges ahead. For this reason, it is fundamental to create quantitative analyses that can answer genuine research questions, therefore allowing a continuous interaction between humanists and computer scientists and using feedback from humanist researchers to create a more suitable system for them. Thus, computer scientists and humanist researchers have more and more interests in common. Although they now share many necessary methodologies, they are still oriented towards different goals. But the digital transformation that is taking place is forcing this confluence, which will increasingly enable this combination. If we want to understand the system in its entirety, we must accept that we lose something. We always pay a price for theoretical knowledge. Reality is infinitely rich; concepts are abstract, they are poor. But it is precisely this poverty that allows us to handle them, and thus to know them. That is why, in reality, "less is more" [3]. This trade-off has characterized our work.

3. Using Digital Tools for Humanities Researchers

While computer scientists and humanist researchers may differ in many ways, they may share certain methodologies. Where computing and the humanities disciplines intersect, DH are created. The use of digital tools for research can greatly simplify manual approaches by allowing more efficient workflows. Qualitative research is a field that includes different analytical strategies, and many researchers use specific programs to assist them in the analysis of qualitative data. Therefore, software vendors are increasingly trying to meet a growing demand for common analytical needs by including more functionalities within their products. Increasingly, researchers deal with a large amount of data gathered through texts, interviews, and observations when conducting qualitative research. This implies a rising need for methodologies to cope with the difficulty of condensing all this information into a research-driven response. The software tools developed to support researchers in these tasks therefore aim to allow researchers to focus on the real goal by saving time and energy, those goals being to analyse data, find the best answers to research questions, and develop theories about them.

Three different approaches can be identified in qualitative analysis: triangulation, mixed methods, and methodological integration. Triangulation in research means using multiple datasets, methods, theories, and/or investigators to address a research question [30,31]. Mixed methods mean the combination of different qualitative and quantitative methods of data collection and data analysis in one empirical research project [32]. Triangulation and mixed methods are oriented towards practical research projects while methodological integration represents a more conceptual and theoretical approach. One of the most commonly used tools in the field of information research, designed to accommodate a mixed-methods approach, is MaxQDA. It is a high-performance data analysis-oriented program for the systematic management and evaluation of texts, documents, and multimedia data, which is used in the medical, scientific, and economic fields for example. One of its strengths is the triangulation of qualitative data that have been collected at different points and at different times, and which are managed through groups of documents that can be analysed in groups or separately. Various types of triangulation are possible [30,31]:

- Triangulation through the investigator—several researchers can take part in an investigation and analyse the same data or parts of it individually;
- Triangulation of theory—different coding schemes can be created and used simultaneously;
- Methodological triangulation—both qualitative and quantitative methods can be used for data analysis by creating a link between them and coding them. In addition, qualitative data can be transformed into quantitative data, which can also be exported and fed into statistical software.

The types of information that the program can process mean that researchers can gather qualitative data through content analysis, enabling them to draw conclusions about the research objective [13]. Using such tools, it is possible to catalogue and structure a significant quantity of texts, making them easier to understand and giving researchers the possibility of using different classification approaches. This avoids the traditional approach of analysing entire files using highlighters and a large amount of paper. The fundamental operation for researchers analysing texts is to emphasize text segments that have a certain importance and to mark them with annotations, comments, memos, and summaries.

4. The Edumeres Toolbox (EDUcational Media RESearch TOOLBOX)

The digital transformation processes in the humanities [33] laid the groundwork for the Edumeres Toolbox [34] to enable the analysis of heterogeneous digital corpora. The Edumeres Toolbox supports researchers in carrying out interdisciplinary research by analysing the data from different perspectives and using results from other services (such as GEI-Digital (https://gei-digital.gei.de/viewer/index/ (accessed on 3 July 2022).) or Curricula Workstation (https://curricula-workstation.edumeres.net/en/about-curricula-workstation/ (accessed on 3 July 2022))).

The web application Edumeres Toolbox is a collaborative platform that offers users the opportunity to collaborate and use natural language processing tools. The application is developed using Java Enterprise Edition (JEE) technology and consists of a frontend that uses JSF and the PrimeFaces framework. JSF provides tools for the construction of the web interface and for the management of the application's navigation flow; PrimeFaces is a suite of open-source components for the creation of the user interface. There is also a backend written in the Java programming language, which works with data stored in a MariaDB database management system (DBMS) and communicates with the frontend to provide the user with the processed data through the graphical interface. The backend also deals with document processing and all the business logic related to the processing operations to be carried out on the documents. It uses different frameworks that constitute a de facto standard both for natural language processing (NLP) and optical character recognition (OCR) in different languages including Arabic, German, English, and Italian (Apache Tika). The end users of the Edumeres Toolbox application are researchers who require quantity analysis tools capable of analysing and processing large quantities of data. The Edumeres Toolbox is a platform based on the idea of edge computing, which means that at the macroscopic level it appears as a system with a distributed and decentralized architecture where each researcher can share resources and texts as well as implemented tools that process data and can be transmitted to other related resources.

In our previous paper we presented the architecture of the Edumeres Toolbox [34] which is based on a broker gateway that orchestrates various functions and modules on the network. In Figure 1 the new toolbox architecture has been designed to meet a more modern design and improve software security, maintainability, and relatability but, most of all, to meet the needs of researchers. The new Edumeres Toolbox, for example, is capable to extract raw text with different OCR tools, to make the analysis of the content using different tools in the same category (i.e., different NLP engines) and to allow the researchers to compare results.

The most important functionalities for researchers are:

- the ability to identify, provide, and manage resources that are connected to the shared network;
- the communication between modules and services;
- the ability to integrate external applications and complex modules;
- the ability to allow network notifications and provide synchronous and asynchronous communication between modules and external applications.

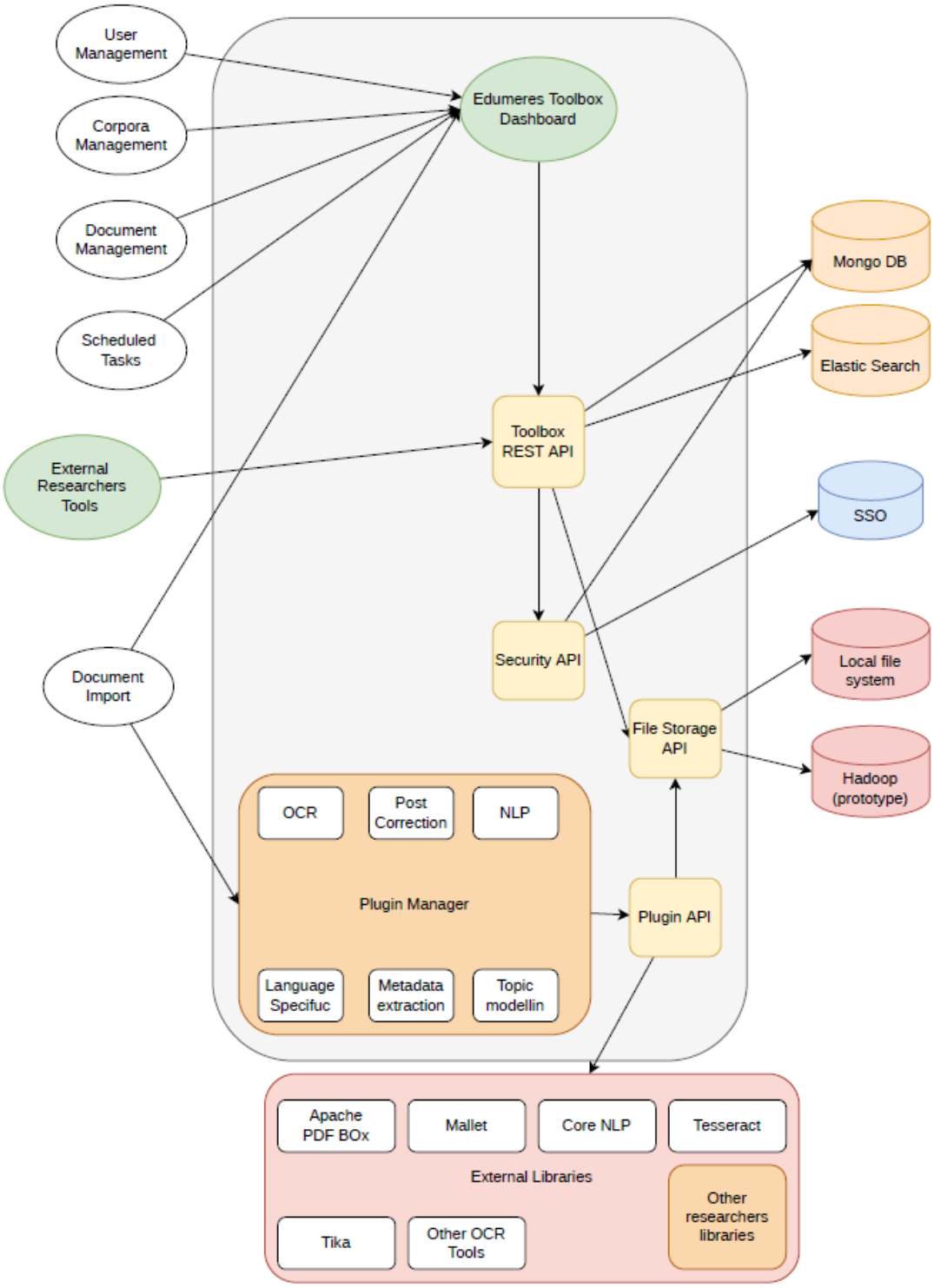


Figure 1. Edumeres Toolbox architecture.

The user interacts with the system through the dashboard and accesses personalized information using the different APIs that have been implemented. The system interacts with other services and modules through standardized interfaces, which have been implemented as modules, as shown in Figure 2.

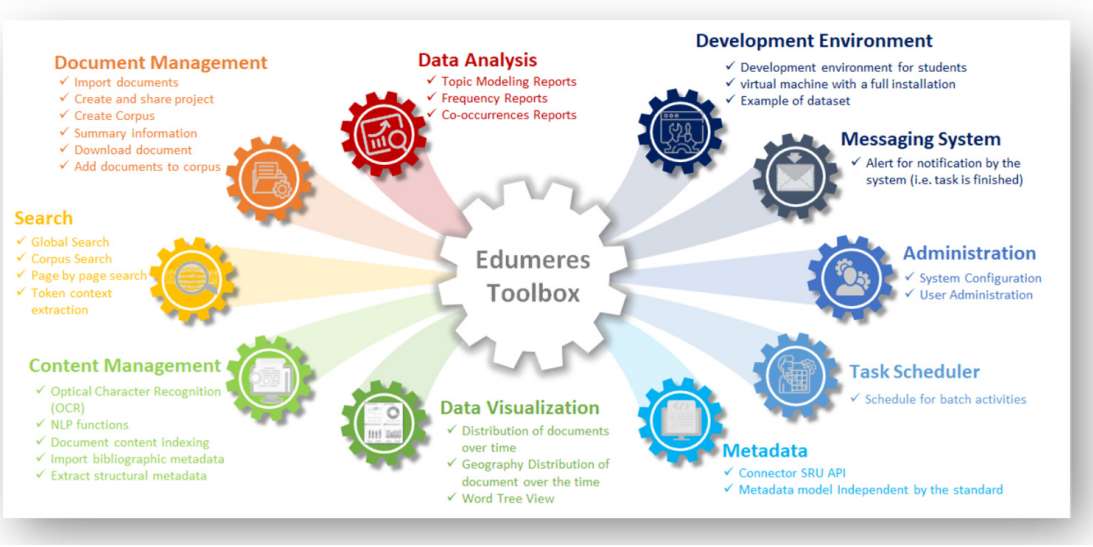


Figure 2. Functionalities of the Edumeres Toolbox.

The main functionalities of the Edumeres Toolbox have been realized as modules and grouped by clusters. The Document Management cluster includes modules that allow users to manage digital documents. Hereby users can create and share projects or corpora, as well as importing and downloading the documents included. The Content Management cluster provides the possibility to extract content from PDF files, start OCR Processing, use NLP functionalities, and extract structural metadata. The SRU connector allows users to automatically import bibliographic metadata, e.g., from the GEI-Digital libraries. Four different search modules (global search, corpus search, page by page search, and the token context extraction) have been implemented, as well as three main text analysis reports, which can be used for more detailed research. Three data visualization functionalities (the time distribution, the geographical distribution, and the Word Tree View) allow users to access data in an interactive way. On the right side of Figure 2, all the administrative and backend functionality clusters are summarized.

When a user wants to create a report, he/she can select the corresponding functionalities. There are some functionalities that are used only once for the corpora, such as the Document and Content or the Meta-data Management, while other functionalities allow users to create reports to meet their need to answer certain research questions. For example, when one user wants to create a frequency or co-occurrence report, they might wish to use the topic modelling functionality. Here, he/she wants to obtain the topics from these reports for quantitative analyses and filter the results, for example, on the basis of the metadata. When a user wants to start his/her research, he/she decides which documents have to be analysed to answer his/her question. He/she then enters his/her account in the Edumeres Toolbox, creates a project, and, for example, shares the documents or corpora with a group of researchers with whom he/she wants to collaborate. Subsequently, he/she imports the documents of interest, and the system performs the Document Management operations, which also include batch processes to extract structural metadata information from the entered documents.

Figure 3 shows a typical simplified abstract workflow, which is enriched with user-independent processes that the system uses to enable the production of reports, providing quantitative and qualitative analyses to respond to user queries. Figure 3 includes the main processes and functionalities, which broadly explain what happens when a report is requested from the system. For example, we have highlighted the modules with the same colours as the functionalities presented in Figure 2. Document Management processes are grouped in orange, Content Management processes are presented in light green, Search processes in yellow, Data Analysis processes in red, and Data Visualization processes in dark green. We have also included the possibility of repeating the operations (partially

or completely), which allows a personalization of the user activities, e.g., to import a new document, to update metadata, or to create several reports using the same corpus.

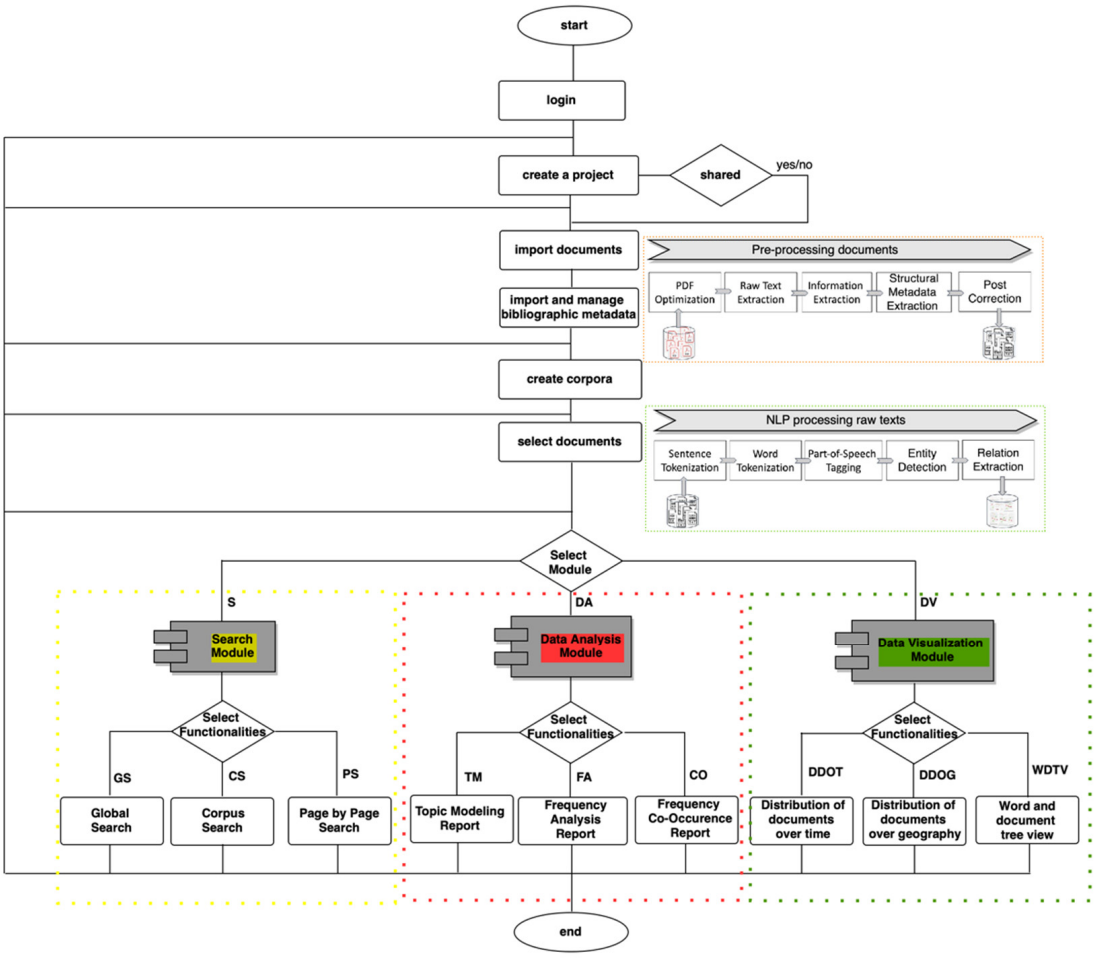


Figure 3. Abstraction of a simplified flowchart and related process flow diagram (light grey) and main module (dark grey) of some use cases to generate reports in the Edumeres Toolbox.

5. Case Study: Analysis of a Collection of Arabic Textbooks

In this work, digital tools are intended to quickly and effectively help humanist researchers in the analysis of large corpora, enabling them to answer research questions and find patterns across large collections of texts by using powerful methods for qualitative and quantitative research.

For instance, to analyse a large collection of Arabic textbooks (around 50 to 100 textbooks) for general education subjects at all school levels, issued over two academic years, the digital tools in the Edumeres Toolbox were used to carry out quantitative and qualitative textual analysis.

The Edumeres Toolbox enabled researchers to create their own projects, insert/import/create metadata, analyse data from different perspectives, compare a group of documents or corpora across multiple criteria and variables (subjects, grades, years, regions, or versions), analyse and visualize extracted information or results in many data visualization types, such as frequency charts and word trees, and export results to Excel or HTML format.

The success of a tool in supporting researchers depends predominantly on how easy the implemented tool is to use and whether it truly gives new access to the needed textual analysis. The Edumeres Toolbox can be used by any researcher to independently carry out different steps: the different assumptions that they may make, the different questions that they may ask, and the different patterns that they may find. All these steps affect the analysis process.

The steps involved in our analysis of a large collection of Arabic textbooks using the Edumeres Toolbox are presented in sequential order in the following pages, from data collection to analysis to interpretation of results.

In the presented use case, we show how the research project on Arabic textbooks is developed using the Edumeres Toolbox. Hereby, the aim of the project was to examine content in textbooks that addressed the promotion of peace, and the peaceful coexistence of two or more groups regardless of their different beliefs, traditions, or cultural heritage, as well as elements addressing tolerance.

As a first step, data were collected. Researchers could create projects/corpora/document collections by uploading/importing them from external resources. The upload file selector allows one or more files to be selected by using the Ctrl and Shift keys.

Researchers can give their project a name and a description (see Figure 4) and control its visibility. Projects are private by default, but it is possible to share them and make them visible to other persons to allow collaborative activity. It is also possible to add categories to a project.

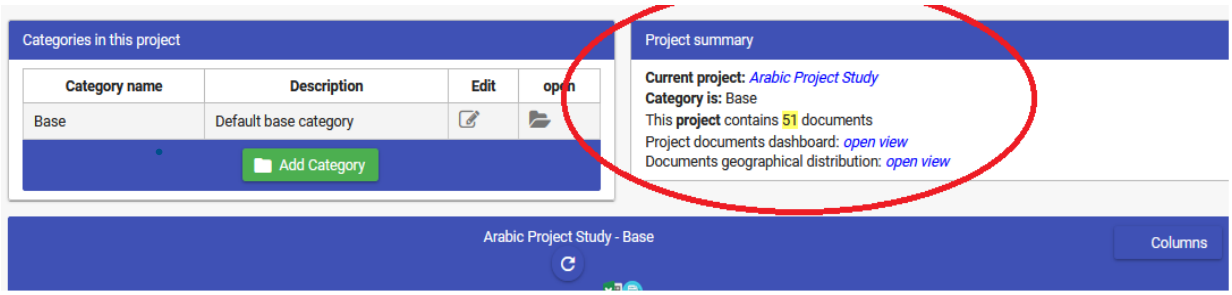


Figure 4. Project summary dashboard.

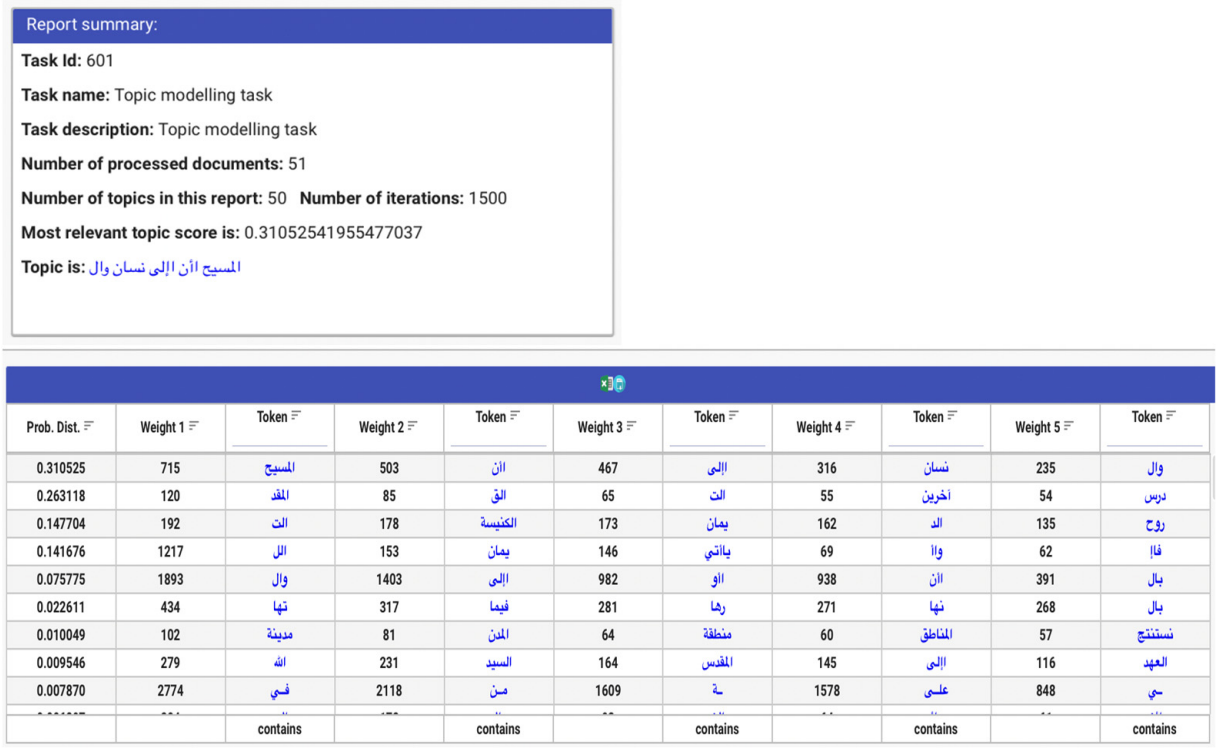
The number of documents in each project may vary from very few to a great number of documents. Before importing the documents themselves, it is necessary to enter or import their metadata. The Edumeres Toolbox includes a component capable of importing metadata automatically from digitized external sources. Metadata include, for instance, the title of the textbook, authors, number of pages, edition, publisher, country, etc.

A summary of the project, or any available shareable project, is displayed on the dashboard when the Edumeres Toolbox is started. The summary contains the name of the current project and its category, specifies how many documents the project includes, and indicates the type and geographical distribution of documents.

Researchers can convert different types of documents into editable and searchable data by using the Tesseract optical character recognition (OCR) engine. This step is very important to obtain accurate results in frequency analysis reports. However, manual corrections are also generally required, especially when converting Arabic language documents as the linguistic characteristics of the Arabic text impose many challenges. The shape of an Arabic character may significantly vary, and this variation depends on the position of the character within a word and its adjacent characters [31]. The fact that Arabic text is written from right to left, is cursive in nature, and can be written in a variety of fonts also means that after the OCR processing, words may be hyphenated or overlap, rendering the document illegible. Arabic words also have diacritics that affect their pronunciation and/or their meaning. The linguistic and grammatical aspects are also a challenge; in the Arabic language, a prefix or suffix could change the whole meaning of a word or make it impossible for a digital tool to recognize the word.

Therefore, in order to improve our frequency analysis results, we propose the creation of black and white lists. This will be addressed in detail under the section on frequency analysis. Returning to the steps carried out by researchers within the Toolbox, we will discuss the natural language processing (NLP) activities such as topic modelling, which serves to detect word and phrase patterns within a set of documents, and the automatic

clustering of word groups and similar expressions that best characterize these activities. These processes help to reveal the abstract "topics" that occur in a collection of documents by determining the probability distribution of topics throughout the text (see Figure 5). Researchers can change the parameters based on their analysis needs and apply filters to group the metadata to improve their research results. The specific aims of the individual analysis will dictate the choices made by the researcher throughout the different steps.



Report summary:

Task Id: 601

Task name: Topic modelling task

Task description: Topic modelling task

Number of processed documents: 51

Number of topics in this report: 50 **Number of iterations:** 1500

Most relevant topic score is: 0.31052541955477037

Topic is: المسيح ان الى نسان وال

Prob. Dist.	Weight 1	Token	Weight 2	Token	Weight 3	Token	Weight 4	Token	Weight 5	Token
0.310525	715	المسيح	503	ان	467	الى	316	نسان	235	وال
0.263118	120	المقد	85	الق	65	الت	55	آخرين	54	درس
0.147704	192	الت	178	الكنيسة	173	يمان	162	الد	135	روح
0.141676	1217	الل	153	يمان	146	ياتي	69	وإ	62	فإ
0.075775	1893	وال	1403	الى	982	او	938	ان	391	بال
0.022611	434	تها	317	فيما	281	رها	271	نها	268	بال
0.010049	102	مدينة	81	المدن	64	منطقة	60	المناطق	57	نستنتج
0.009546	279	الله	231	السيد	164	المقدس	145	إلى	116	العهد
0.007870	2774	في	2118	من	1609	ة	1578	على	848	سي
		contains		contains		contains		contains		contains

Figure 5. Topic modelling.

In the frequency analysis, the Edumeres Toolbox shows the overall word frequencies for the entire corpus as well as the information about how word frequencies are spread out across the documents within the corpus. The researcher can control the process and refine the results based on his/her needs by applying filters to omit words that may affect the results and by adding similar terms to exclusion lists, for example. Excluded words are those that are defined as irrelevant to the analysis or those listed in the abovementioned black list. This functionality can be used, for example, to remove words that have been incorrectly extracted by OCR to make sure the data in the frequency reports are clean. In addition, users can use "stop-word" lists specific to each language. These include so-called "function words" that do not carry much meaning, such as determiners, prepositions, definite and indefinite articles, conjunctions, or numerals (the, a, in, is, at, which, on, to, from, etc.). Stop-word lists can be manually curated for each language, and these words are filtered out before the data (text) undergoes natural language processing, or after.

In our research study, for example, we wanted to examine the number of times words such as "peace", "culture", "religion", etc., appeared in each textbook and to compare the results according to metadata categories such as subject, grade, and year (see Figure 6).

To do this, we selected the stop-word list for the Arabic language, removed/added prefixes and suffixes as needed, and performed the frequency analysis. By interpreting the results, we were able to draw conclusions, answer our research questions, and prepare a report detailing the findings, interpretations, and conclusions.

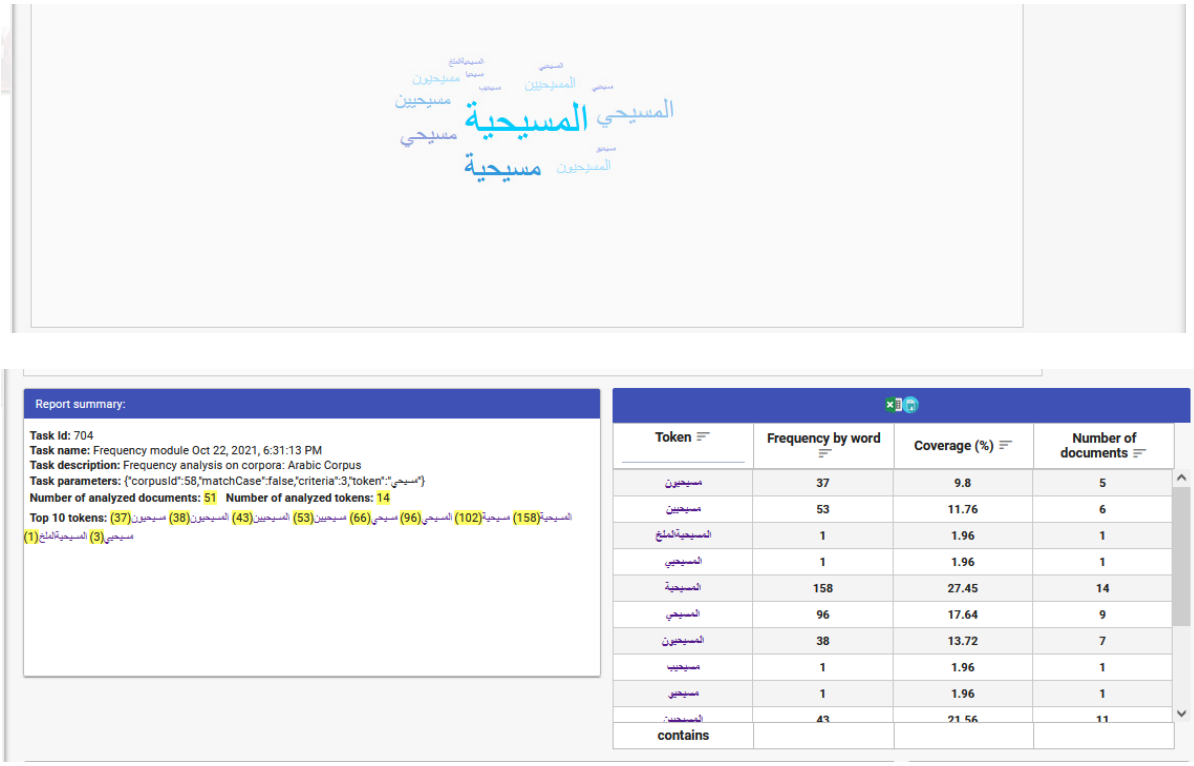


Figure 6. Frequency analysis report for the word "Christian" or "masīḥī" "مسيحي" to find out, for example, how many times it is referred to as the "other" with his/her religion.

For instance, we wanted to know how many times the word "peace" or "سلام" (*salām*) appears in each textbook, how the use of the term changes from year to year or from grade to grade, and in which context it appears. When we selected the word on its own, without excluding any similar words, the tool also counted the word 'إسلا' (*Islām*), which means "Islam" (the religion), making our research inaccurate. The fact that the two words are very similar made it hard for the machine to recognize the difference. Therefore, by adding the word 'إسلام' (*Islām*) to the black list, we ensured it would not appear in future analyses and obtained more accurate results. Patterns could, for example, be found in changes in the frequency of the word in new editions, leading to different conclusions such as promotion of peace education, cultural tolerance or intolerance, respect for other beliefs, etc.

Another example is the frequency analysis of words such as "Christian" or "Islam"; how many times do they refer to the "other" with his/her religion? Are some religions being neglected or distorted in these textbooks? Is reference to religion made especially in the context of cultural differences? Such analyses can lead the researcher to draw conclusions and answer research questions.

The findings were generated in the form of graphs and charts and exported in Excel format (Figure 7). Clicking on each token enables researchers to see the document(s) in which each word appears as well as its frequency. By then clicking on the document title, the word appears in its context alongside information regarding the syntactic structure or context-free grammar in which it is used. All this enables the researcher to closely interpret the results. In the report preferences, researchers can use the filter function to find out, for example, the frequency with which a certain word occurs in a specific educational level or year, or to group the results based on the metadata.

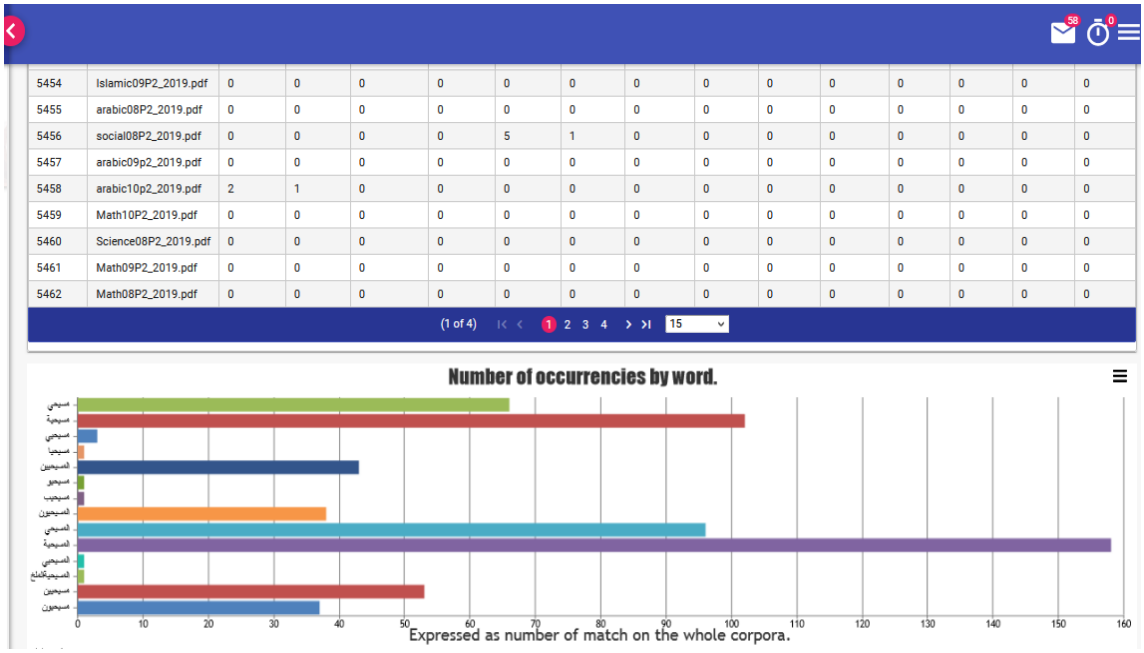


Figure 7. Graph displaying the occurrences of the word "Christian" or "مسيحي" (masīḥī) in each subject.

Although objective figures can be obtained from quantitative analysis, which can lead to important answers and conclusions, human intervention is also needed in order to interpret the results. For example, in the frequency analysis of the word "war" or حرب(*ḥarb*) in Arabic, the tool counted all words that contain these three letters. However, these words (see highlighted words in Figure 8) do not all mean "war" (*ḥarb*): 'حرباء(*ḥarbā'*) means "chameleon", for example. Therefore, by excluding these words or by adding them to the "blacklist", a researcher can get reliable results and draw accurate conclusions. Without the exclusion of irrelevant words in the example given, the frequent use of the word "war" in the textbooks could lead to inaccurate conclusions of an escalation in conflict, in addition to many other interpretations that could have serious consequences due to the sensitivity of the topic.

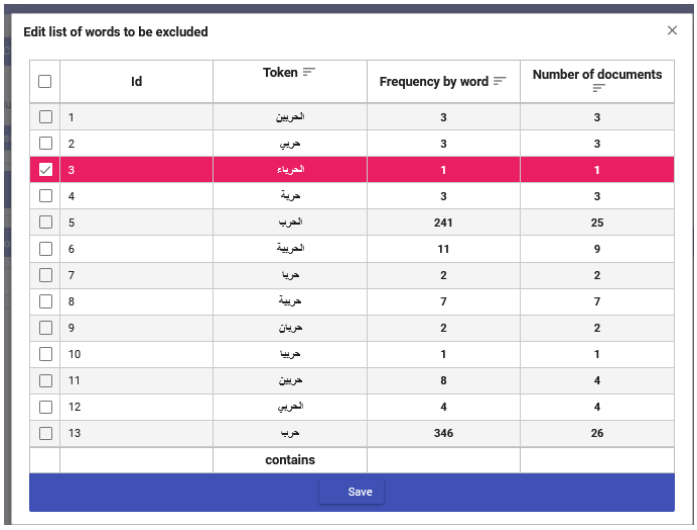


Figure 8. Two lists of selected words to be excluded from the frequency report as they differ in meaning from the search term.

6. Conclusions

The Edumeres Toolbox allows researchers to import, analyse, and compare data in ways that produce new information. Textual analysis using the Edumeres Toolbox is convenient and functional, and it does not require in-depth computer science skills. However, the quality of the OCR or digitized texts, especially in the Arabic language, has proven to be a major challenge, and one that must be addressed. Putting a text through an OCR engine does not necessarily guarantee that it will be interpretable by the Edumeres tools. Thus, additional steps, such as human intervention to manually identify words and characters, may be required to improve the results.

Since its development in 2018, the Edumeres Toolbox has demonstrated its effectiveness as a means of digital textual analysis for humanist researchers, as it allows them not only to answer their research questions but also to discover new potential questions and patterns. Humanist researchers have proven to be the main acting party in this process, as their contributions, ideas, and assessment of tool usability and functionality are crucial at all development stages. Therefore, their feedback enables the Toolbox to improve its performance and move closer to achieving its objectives (Please review the objectives of the Edumeres Toolbox stated at http://www.gei.de/en/flagships/edumeres-toolbox.html (accessed on 3 July 2022).).

Author Contributions: Conceptualization, F.F.; Formal analysis, F.F.; Methodology, F.F.; Supervision, R.S. and E.W.D.L.; Validation, F.F.; Visualization, F.F.; Writing—original draft, F.F.; Writing—review & editing, B.G. and E.W.D.L. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Data Availability Statement: Not applicable.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Hawthorn, J. *A Glossary of Contemporary Literary Theory*; Oxford University Press: Oxford, UK, 2000.
2. Piper, A. Novel devotions: Conversional reading, computational modeling, and the modern novel. *New Lit. Hist.* **2015**, *46*, 63–98. [CrossRef]
3. Moretti, F. Conjectures on World Literature, «The New Left Review», II, 1 2000. Available online: https://newleftreview.org/issues/ii1/articles/franco-moretti-conjectures-on-world-literature (accessed on 7 April 2022).
4. Moretti, F. *Graphs, Maps, Trees: Abstract Models for a Literary History*; Verso: London, UK, 2005.
5. Moretti, F. *Distant Reading*; Verso: London, UK, 2013.
6. Beals, M. TEI for Close Reading: Can It Work for History? Available online: http://tinyurl.com/nvdndsb (accessed on 1 September 2015).
7. Bradley, A.J. Violence and the Digital Humanities Text as Pharmakon. In Proceedings of the Digital Humanities 2012, DH 2012, Conference Abstracts, Hamburg, Germany, 16–22 July 2012; Meister, J.C., Ed.; Hamburg University Press: Hamburg, Germany, 2012.
8. Coles, K.; Lein, J.G. Solitary Mind, Collaborative Mind: Close Reading and Interdisciplinary Research. In Proceedings of the Digital Humanities 2013, DH 2013, Conference Abstracts, Lincoln, NE, USA, 16–19 July 2013; pp. 150–152.
9. Schuster, K.; Dunn, S. (Eds.) *Routledge International Handbook of Research Methods in Digital Humanities*, 1st ed.; Routledge: London, UK, 2020. [CrossRef]
10. Jnicke, S.; Franzini, G.; Cheema, M.F.; Scheuermann, G. On Close and Distant Reading in Digital Humanities: A Survey and Future Challenges. A State-of-the-Art (STAR) Report. Available online: https://www.informatik.uni-leipzig.de/~{}stjaenicke/Survey.pdf (accessed on 3 July 2022).
11. Shneiderman, B. The Eyes Have It: A Task by Data Type Taxonomy for Information Visualizations. In Proceedings of the IEEE Symposium on Visual Languages, Darmstadt, Germany, 5–9 September 1996; pp. 336–343.
12. De Luca, E.W.; Fallucchi, F.; Ligi, A.; Tarquini, M. A Research Toolbox: A Complete Suite for Analysis in Digital Humanities. In *Communications in Computer and Information Science*; Springer: Cham, Switzerland, 2019; 1057 CCIS; pp. 385–397.
13. Gay, L.; Mills, G.; Airasian, P. *Educational Research: Competencies for Analysis and Application*, 9th ed.; Lawrance Erlbaum Associates: Puplishers, NJ, USA, 2009.
14. Ascari, M. The Dangers of Distant Reading: Reassessing Moretti's Approach to Literary Genres. *Genre* **2014**, *47*, 1–19. [CrossRef]
15. Ahmed, P.; Al-Ohali, Y. Arabic Character Recognition: Progress and Challenges. *J. King Saud Univ.* **1999**, *12*, 87.
16. Richards, I.A. *Practical Criticism*; K. Paul, Trench, Trubner: London, UK, 1929.

17. Saito, S.; Ohono, S.; Inaba, M. A Platform for Cultural Information Visualization Using Schematic Expressions of Cube. Available online: https://dh2010.cch.kcl.ac.uk/academic-programme/abstracts/papers/pdf/ab-796.pdf (accessed on 3 July 2022).
18. Porsdam, H. Digital Humanities: On Finding the Proper Balance between Qualitative and Quantitative Ways of Doing Research in the Humanities. *Digit. Humanit. Q.* **2013**, *7*, 1.
19. Wolski, U. *The History of the Development and Propagation of QDA Software*; University of Northampton: Northampton, UK, 2018.
20. Bosch, R. Qualitative research in the digital humanities. *Editor. Kwalon* **2016**, *21*, 1. [CrossRef]
21. Glaser, B.G.; Strauss, A.L. *The Discovery of Grounded Theory: Strategies for Qualitative Research*; Aldine Publishing: Chicago, IL, USA, 1967.
22. Weber, M. Die 'Objektivität' sozialwissenschaftlicher und sozialpolitischer Erkenntnis. *Arch. Für Soz. Und Soz.* **1904**, *19*, 22–87.
23. Znaniecki, F. *The Method of Sociology*; Farrar & Rinehart: New York, NY, USA, 1934.
24. Glaser, B.G. *Theoretical Sensitivity: Advances in the Methodology of Grounded Theory*; Sociology Press: Mill Valley, CA, USA, 1978.
25. Bosch, R. Pragmatism and the practical relevance of truth. *Found. Sci.* **2007**, *12*, 189–201. [CrossRef]
26. Bosch, R. *Wetenschapsfilosofie Voor Kwalitatief Onderzoek*; Boom Lemma Uitgever: The Hague, The Netherlands, 2012.
27. Bosch, R. Power: A conceptual analysis. The Hague: Eleven International Publishing. In *The SAGE Handbook of Grounded Theory*; Bryant, A., Charmaz, K., Eds.; SAGE: Los Angeles, CA, USA, 2007; *in press*.
28. Kahneman, D. *Thinking, Fast and Slow*; Farrar, Straus & Giroux: New York, NY, USA, 2011.
29. Jockers, M.L. *Macroanalysis: Digital Methods and Literary History*; University of Illinois Press: Chicago, IL, USA, 2013.
30. Denzin, N.K. *Sociological Methods*; McGraw-Hill: New York, NY, USA, 1978.
31. Patton, M.Q. Enhancing the quality and credibility of qualitative analysis. *HSR Health Serv. Res.* **1999**, *34*, 1189–1208. [PubMed]
32. Burzan, N. Review: Udo Kuckartz. Mixed Methods. Methodologie, Forschungsdesigns und Analyseverfahren [Mixed Methods. Methodology, Research Designs and Methods of Data Analysis]. Forum Qualitative Sozialforschung/Forum: Qualitative Social Research. *FQS* **2014**, *16*. [CrossRef]
33. De Luca, E.W.; Spielhaus, R. Digital Transformation of Research Processes in the Humanities. In *Communications in Computer and Information Science*; Springer: Cham, Switzerland, 2019; pp. 343–353.
34. De Luca, E.W.; Fallucchi, F.; Nobili, C. Edumeres Toolbox: Functional, Technical, Architectural Analysis. In *Communications in Computer and Information Science*; Springer: Cham, Switzerland, 2022; 1537 CCIS; pp. 212–223.