Computational Mining on IBICT BDTD's Thesis and Dissertation Metadata for Supporting Social Science Research*

Rodrigo R. Filho[1], Elismênnia A. Oliveira[2], Jordão H. Nunes[2], Marcelo A. Inuzuka[1], Hugo A. D. do Nascimento[1]

[1]Instituto de Informática, [2]Faculdade de Ciências Sociais
Universidade Federal de Goiás (UFG)
Av.Esperança sn, Câmpus Samambaia – 74690-900 – Goiânia – GO – Brazil

{rodrigorfh,mennalis}@gmail.com, {jordao_fcs,marceloakira,hadn}@ufg.br

Abstract. *The Brazilian Digital Library of Thesis and Dissertations (BDTD) provides essential data to support many social sciences investigations. Nevertheless, there is still a lack of computation tools tailored for helping extract and analyze the necessary information from the BDTD library. In this paper, we discuss the development of computational solutions to answer questions from a particular social sciences research using metadata from BDTD. The solutions involve the integration of data processing and presentation techniques, such as string-processing algorithms, knowledge graphs and information visualizations. All programming codes implemented at the scope of the project are available for helping other researchers. The paper also highlights the importance of having researchers from Social Science and Computer Science working together, what motivates future collaborations in these areas.*

1. Introduction

The expansion of computer networks, storage and data collection capacity generates a complex and heterogeneous volume of data, coming from different sources, whose controls are distributed and decentralized, constituting a set called big data [McAfee and Brynjolfsson 2012]. An example of this type of set is the Brazilian Digital Library of Thesis and Dissertations (BDTD), which databases are locally updated and maintained in each educational institution and have their metadata further harvested and aggregated by the Brazilian Institute of Information in Science and Technology (IBICT). The BDTD and other similar databases containing information about scientific production and the academic trajectories of researchers, such as the Brazilian CNPq's Lattes Platform and the Sucupira Platform, allow the development of studies and applications in distinct scientific areas, such as computing and information and social sciences. Particularly relevant here are the studies on data mining and knowledge generation, which are strongly dependent on computational data processing techniques to generate content or knowledge [Kambatla et al. 2014, Zhang et al. 2014]. In fact, several scientific papers addresses the usage of computational methods for mining and analyzing data from the Lattes Platform [Mena-Chalco and Cesar Junior 2009],[Alves et al. 2011]. However, in contrast to what has been produced about Lattes, there is still a lack

*This study was financed in part by the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brasil (CAPES) - Finance Code 001. The authors also thank IBICT for providing the requested dataset.

of works that present computational tools for supporting scientific investigation of the BDTD, in particular in the information and social sciences fields. The majority of that research, such as the works of Campello et al. [Campello et al. 2007], Bucher-Maluschke et al. [Bucher-Maluschke et al. 2019], Silva et al. [Silva et al. 2018], Hayashi [Hayashi et al. 2007] and Santos Júnior and Real [Santos Junior and Real 2017] used only small data samples and focused on the complete analysis of the selected texts instead of exploring the metadata. In general, those works involved simple searches on the documents with keywords and the analysis of data using spreadsheets. There are broader works such as the ones of Costa and Rodrigues [Costa and Rodrigues 2019], and Vitta et al. [Vitta et al. 2018], which addressed issues such as interoperability between databases, and the processing and visualization of BDTD data, but they did not infer categories of analysis neither described in detail the used tools.

The current paper is an effort to fill the above gap. It provides an insight into a computation-based study of BDTD's metadata for supporting a social science research project. Three questions regarding the discussion of some specific social themes by works in the BDTD database are considered. The usage of computational techniques and tools – for example, string processing algorithms and information visualization methods – for helping answer those questions are presented. A special attention is given to the usage of knowledge graphs for modeling and analyzing some of the pieces of information extracted from the BDTD. All programming codes implemented at the scope of the project are publicly available, in order to further help other researchers.

The remainder of this paper is organized as follows. Section 2 details the social sciences research project that motivated this work. Section 3 describes the approach employed for answering the questions raised in the research project. Section 4 presents our conclusions and suggests ideas for future studies.

2. The Social Science Research Project

This paper focuses on a research project for understanding the intersection between gender and race categories in academic publications in Brazil. The overall goal of the project is to map the rising of that intersection over the years, its geographic distribution in the country and the main related themes discussed by academic authors according to their sex. An initial set of themes (such as "racism", "Africa", "male", "female", etc.) was defined and several research questions were formulated to be answered as part of the project. Due to space limitation, only a group of three questions are discussed here, as listed below:

(Q1) What are the main topics of the master dissertations and doctorate thesis for each theme? How related are those topics to the academic research done in every institution and geopolitical region?
(Q2) How are the works distributed according to their knowledge area, year of publication and institution?
(Q3) How are the student-advisor matrices by theme, year and geopolitical region?

The choice for using BDTD in order to answer these questions was due to its national representativeness, the possibility of extracting data from it in CSV and JSON formats, and the speed in which IBCIT could reply to researchers' requests. The BDTD was created by IBICT in 2002 with the aim of promoting and facilitating access to graduate studies. Currently, with over six hundred thousand documents from 118 institutions,

it is the largest online library of its kind in the world, and the only one in Brazil with a public user interface that allows direct bulky data extraction from its own platform [1].

Data extraction from the BDTD Web system can be done by specifying author names, work titles and subject keywords in a simple graphical search interface, or by an advanced search tool based on keywords and filters (the filters can be set to select documents according to their type – e.g master or doctorate thesis –, idiom and year of publication). Eighteen keywords were defined for covering all themes. These keywords are roots of Portuguese terms that define the themes, followed by the wildcard character (*). Examples are "mascul*", which covers the terms "masculino", "masculinização" and "masculinidade"; and "raci*", which includes "racial", "raciais", "racismo" and "racialização". Each root was used in a single search, resulting in a list of metadata entries to be downloaded as a CSV file. However, there was a maximum limit of 1000 entries that could be exported for each search. This allowed the extraction of only seven of the eighteen desired contents. A formal request was issued to IBICT by email, which provided later the remaining eleven CSV files.

Table 1 shows the fields for one entry of a table extracted from the BDTD. It is important to note that not all cells of the CSV tables are filled, since some details about the academic works may be missing. In addition, there are duplicated data (the same information appearing twice, by mistake, in two columns) and very compressed information in some columns (such as the "contributors" column, which has between three and eight parts with different functions). All of these details demand a preprocessing phase, which, by itself, makes purely manual work unfeasible.

Next, we present the approach created in the scope of the current project for support answering the above-mentioned research questions.

3. Investigation Approach

A computational pipeline with several tasks was designed for this work. It is illustrated in Figure 1 and consists of the six main elements, described below in the order in which they are used:

1. **Raw CSV**. These are the original CSV files with metadata extracted from BDTD. As mentioned in the previous section, we have 18 raw CSV files, one for each research theme.
2. **Preprocessing**. Preprocessing routines extract and reorganize more detailed information from the raw CSVs. They write their results in new CSV files or in a knowledge graph structure.
3. **Modified CSVs**. These are new CSV files that contain useful and well-formatted data for further analysis.
4. **Knowledge Graph (KG)**. This is a graph-based representation of knowledge extracted from the CSV files, that is suitable for more advanced information processing and for data-relation visualization.

[1] We note that there are other large Brazilian digital libraries that hold scientific and academic information, such as the previously mentioned Lattes Curriculum platform, the CAPES Group of Theses and Dissertations and the Brazilian Portal for Scientific Publication in Open Access (Oasisbr). However, despite the Law on Access to Information, LAI No. 12527/2014, the extraction of data stored in those databases is not a simple and direct process. The extraction methods adopted are in general restricted to very limited views.

Table 1. Most relevant fields of a CSV table extract from BDTD.

id	Id code of the dissertation/thesis
title	Title of the work
abstract_por	Abstract in Portuguese
abstract_eng	Abstract in English
authors	Author's details
contributors	Information about the advisor, co-advisor and referees
subjectsCNPQ	List of categories according to CNPq Classification
subjectsPOR	List of keywords in Portuguese
subjectsENG	List of keywords in English or "NaN"
institutions	Acronym of the institution where the work was done
departments	Academic department of the institution
programs	name of the postgraduated program
types	"masterThesis", "doctorateThesis" or "bachelorThesis"
accesslevel	"openAccess"
publicationDates	Year of publication
urls	URL of the work
languages	Code of the used idiom. E.g.: "por", "eng", "spa" or empty

5. **Question-answering focused tasks**. With all necessary data available in either CSVs or KG formats, we perform tasks that produce highly significant information, mostly presented in the form of visualizations.
6. **Visualizations**. These are charts, diagrams and other types of information visualizations for supporting answering the research questions.

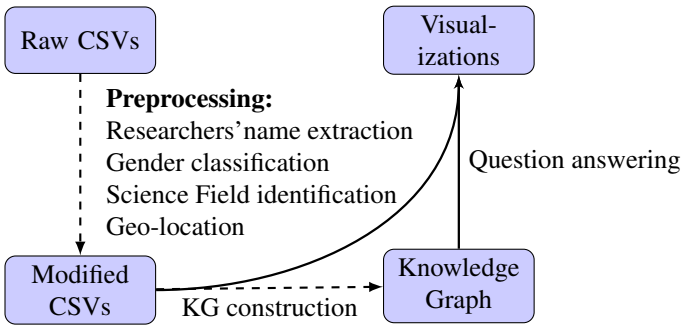

Figure 1. Computational pipeline for helping answer the research questions.

The work was done through a collaboration between social sciences researchers and computer scientists. In order to facilitate communication and knowledge transfer between the investigators (in particular, regarding the computation methods that were designed and/or used), all the data processing steps and programming code developed within the scope of this project were documented in Jupyter notebooks using Google Colaboratory platform. Weekly meetings were organized to discuss the demands, analyze data, present the programming routines developed and validate the generated solutions. In the next subsection, we describe in detail all preprocessing tasks and question answer tasks performed during the project. For each task, one or more programming codes were created and they are publicly available at `https://gitlab.com/ivato/bdtd/eniac`.

3.1. Preprocessing

3.1.1. Researchers' name extraction

The aim of this task is to extract, for each dissertation and thesis, the name of the author, as well as the name of the advisor and the referees (the other members of the evaluation board). These pieces of data are recorded in two columns of the raw CSV files: "Authors" and "Contributors" (the latter with information about the advisor, co-advisor and referees). Many different formats have been observed in these fields: with names and surnames written in direct or reverse order; with or without date of birth; and sometimes with a URL.

The computational solution for such a preprocessing was to use regular expressions in order to identify and extract the necessary information. This choice was guided by the finite amount and similar patterns of the formats that appeared in those columns. In addition, the Python RE package offers routines for identifying and extracting parts of strings using regular expressions, which can be applied easily on the table. In this way, functions were built to separate and normalize the full names of authors, advisors, co-advisors (when they exist) and referees. The functions were assembled in a program that call them for all thematic CSV files, generating, for each one, an extended CSV spreadsheet with the new columns containing the preprocessed data.

3.1.2. Gender classification

Gender identification was performed for authors, advisor, co-advisor and referees using the new column with their names. The Python Gender Guesser [2] library was employed for this purpose, by applying it to the first name of each person.

The Gender Guesser employs the dataset *nam-dict.txt*, which has more than 40,000 first names of people and their most commonly associated gender in several countries, including Brazil. The gender identification routine returns, for a given name, one of the possible results: male, mostly male, female, mostly female, androgynous and not found. The male and female classifications were considered as a definitive answer in the analysis. The androgynous and not-found results were mapped to Not-defined.

However, it was noted that many common names in Brazil were not in the *nam-dict.txt* database, such as: Ademir, Evani, Glaucio, Jacimara, Luiz, Matheus, Mônica, Tâmara, Thaís and Thiago. For the missing names and for all others which classification were not clear in the previous analysis (e.g., mostly male, mostly female and androgynous), a second verification was carried. This verification used a spreadsheet[3] with the probable gender for several proper names in Brazil, built from the IBGE Demographic Census base[4]. Only the names not present in the spreadsheet and which also were marked as androgynous or not-found in the previous search remained unclassified. The remainder was labeled as either male or female.

From a total of 104,873 distinct first names in the CSV files, only 4676 could not

[2]https://pypi.org/project/gender-guesser

[3]`https://brasil.io/dataset/genero-nomes/nomes/`

[4]`https://censo2010.ibge.gov.br/nomes`

be classified (were not found in both databases). They were mostly abbreviated (like "P." for "Paulo") and some others had character encoding problems. The gender information was added to new columns of the extended spreadsheets mentioned in the last section.

3.1.3. Science Field identification

The CSV files have the column "subjectsCNPq" with information that defines the CNPq's Science classification [5] for every dissertation and thesis. This column follows a structured format consisting of a list of sequences of terms. Each sequence may have one or more terms in hierarchical order, going from the most general science field to the most specific one. Sometimes, however, the sequence has only a more specific term, not mentioning the whole hierarchical branch. Examples of such information for two particular documents are "CIENCIAS AGRARIAS::AGRONOMIA::FITOTECNIA | CIENCIAS BIOLOGICAS::BOTANICA::FISIOLOGIA VEGETAL" and "Ecologia|Sensoriamento Remoto". The symbol '|' divides lists, while '::' represents the separation between two consecutive terms in a hierarchical sequence.

The aim of the current preprocessing is to find as many science fields as possible that are related to a dissertation or thesis. For this, a Python code was written for directly collecting CNPq's Science classification from the Web and represent it internally as a JSON hierarchical structure of terms. Then, for every publication, its "subjectsCNPq" attribute was analyzed and the last term of each sequence was taken. Next, these last terms were searched in the JSON structure for recovering their complete CNPq hierarchical paths. Finally, a set with all CNPq paths that were found for that particular work was created and saved in a column of the CSV dataset, for future usage. An important observation is that we used the Levenshtein measure in the search routine, so that, even if a term was mistyped in the BDTD database, we still get the most similar CNPq field.

3.1.4. Institutions' Geo-Location

A preprocessing was necessary in order to find the Geopolitical (Northen, Northeastern, Midwest, Southeast and South) regions in Brazil of the institutions in which the dissertations and thesis were presented. The "Institution" column in the CSV files is the starting point for this task. A total of 113 unique academic institutions were identified using that column. For each one of them, the city of its main campus were manually searched on the Web. Next, the IBGE code of the city was recovered (also in a manual fashion) from the website `https://cidades.ibge.gov.br/`. With all IBGE codes in hand, a routine was implemented to automatically retrieve the name of the mesoregion of all institutions using an IBGE's Web service[6]. As result, a new CSV file was created with the acronym, city, IBGE code and region of all academic institutions. Note that, given the IBGE codes, it is possible to apply openly available Python tools for collecting additional information about the cities, including their geographic coordinates (for instance, see the script at `https://github.com/kelvins/Municipios-Brasileiros`).

[5]CNPq is a Brazilian science research founding agency (`http://www.cnpq.br`).

[6]The Web Service is at: `https://servicodados.ibge.gov.br/api/v1/localidades/municipios/CODE`, where CODE must be replaced by the IBGE code.

3.1.5. Construction of the Knowledge Graph

In order to build a knowledge graph containing some of the information present in the CSV files, we used a graph-based database system, in this case *Neo4j*[7]. When installed, Neo4j provides a Web interface through which it is possible to build, maintain and consult, and even visualize, a general graph. A specific script language, called *cypher*, allows to perform those and many operations on graphs. In the scope of this project, we wrote cypher scripts that read all CSV files and build a single knowledge graph with the following characteristics:

- Each academic work (an entry in a CSV table) was represented by a node called "Publication" containing seven attributes (*Id* as a unique identification code, *title*, *type* (master, PhD or other), *language* for expressing the written language of the work, *publicationData* as a data attribute, and *themes* and *CNPq_class* as a list of terms);
- Each author, advisor, co-advisor and referee was modeled as a node called "Person" with two attributes (*name* and *gender*);
- Each institution to which a master or doctorate work is linked was mapped to a node "Institution", also with two attributes ((*name* and *region*);
- An edge in the graph was created for every type of relation as described below, with the first and latest terms specifying the initial and final nodes, respectively, and the term in brackets representing the relation that defines the edge's type.

  Person [ADVISOR] $\rightarrow$ Publication
  Person [REFEREE] $\rightarrow$ Publication
  Person [AUTHOR] $\rightarrow$ Publication
  Person [COADVISOR] $\rightarrow$ Publication
  Publication [DEPOSITEDBY] $\rightarrow$ Institution

3.2. Question Q1

The aim of this question is to understand what the dissertations and thesis in each one of the eighteen themes (files/datasets) talk about, attesting the recurrence of some of the main discussed topics. Word clouds of type Wordle were created for that goal, one for each theme, by joining texts from the columns *abstract_por* of the related CSV file. Figure 2 shows a word cloud produced for the theme "mascul" referring to the academics works with keywords "masculino", "masculinas" and "masculinidade", among others.

As we can see in Figure 2, the word cloud shows very recurring topics: male and female sexes, patient, women, man, child, gender, time, individual, treatment, population, control, age group. In addition, it contains less recurring keywords such as: culture, social and HIV AIDS. This visualization demonstrates that the academic works are within the scope of what is expected in humanities studies on masculinity. It also shows that other scientific areas are present such as Applied Social Sciences, Health and Biology.

For the second part of the question, the interest is to get details about how the topics are pervasive in each academic institution and which topics are mostly approached in each geopolitical regions. Since there are much data to be visualized in low level, we opted here for using an interactive treemap visualization based on the 3d.js library

[7] https://neo4j.com/

Figure 2. Main topics discussed in the "Mascul" CSV file.

embedded into a HTML file. A Python code was developed for processing the modified CSVs and creating and saving a JSON tree description of the amount of dissertations and thesis, grouped hierarchically by region, institution, theme and main topics. By opening the HTML file, the treemap is shown and the user can drill in and out through the data levels. Figure 3 illustrates such a treemap.

Figure 3. Treemap for exploring the presence of the main discussed topics.

3.3. Question Q2

Question Q2 is very similar to the previous one. The main difference is that, instead of analyzing the academic works from the point of view of their discussed topics, it aims at elucidating the relationship between the major areas of knowledge. For instance, which areas have a greater and a lesser presence of publications for each considered themes. There is also a concern with the understanding of the temporal changes in research focus.

Therefore, for helping answer this question, a treemap was used again. The hierarchical structure of the previous treemap was employed, except that the studied topics (described by keywords) were replaced by the preprocessed CNPq fields and the column *PublicationDates*.

By interpreting the word cloud created for the previous question together with this new interactive treemap (Figure 3), it is possible, for instance, to realize that most studies referring masculinities are allocated in the southeastern and southern regions of Brazil, and that the areas of Health and Applied Social Sciences have studied this theme for a long period of time. Several other conclusions are possible by exploring these visualizations.

3.4. Question Q3

The aim of Question Q3 is to bring light to the connection between students and advisors, possibly identifying some strong links around particular themes and regions. For adequately answering this question, we opted for using the KG previously built, but enhancing it with an extra type of edge that makes the advisor-student relationship more explicit. The Neo4j was also employed.

A Cypher script was implemented in Neo4J to add new types of edges called "Advisor of" and "Co-advisor of" to the graph. We note that this relationship already exists in KG, albeit implicitly, through people who are connected to a node of type "publication".

The Neo4j graphical interface was then used to specify a search query and view the results as a graph drawing in a dynamic fashion. For the current task, the query resulted in a student–(co-)advisor graph with nodes that represent the people and directed edges that describe the advisor's relationship between them. Figure 4 shows part of this graph drawing for the theme "mascul" with ten advisors and their former students, two of them supervising a same student. From this type of visualization, academic networks and concentration patterns can be realized. Some advisors may have a much larger number of students around them, which suggests that they can be principal researchers in that particular area of knowledge. Supervision paths in which alumni became advisors for new students also demonstrate the influence of some researchers in advancing an area. Extra rules can be added to the query to filter nodes and edges related to specific years, geopolitical regions and institutions.

4. General Remarks and Conclusion

Without the aid of computational tools for helping process and analyze a significant amount of data, many social sciences studies are impractical nowadays. One can easily find a variety of tools, both proprietary and opensource, for supporting social science researchers. These include statistical-analysis software (like SPSS and R), text-semantic analysis tools (e.g, NVivo, Atlas-ti, etc.), general data mining and visualization tools (like Tableau, PowerBI and Metabase), graph-based data visualization software (like Gephi, YEd and Grahviz) and so on. Nevertheless, no single software provides a complete solution to all needs. In the best case, they can be used for supporting only part of the work. In this realm, the development of specific computational routines through code programming is still an essential resource for dealing with most open problems. Computer programming certainly demands skills that not all social science researchers have. But this can be compensate in a fruitful way through a collaboration with computer scientists, as demonstrated in the present work.

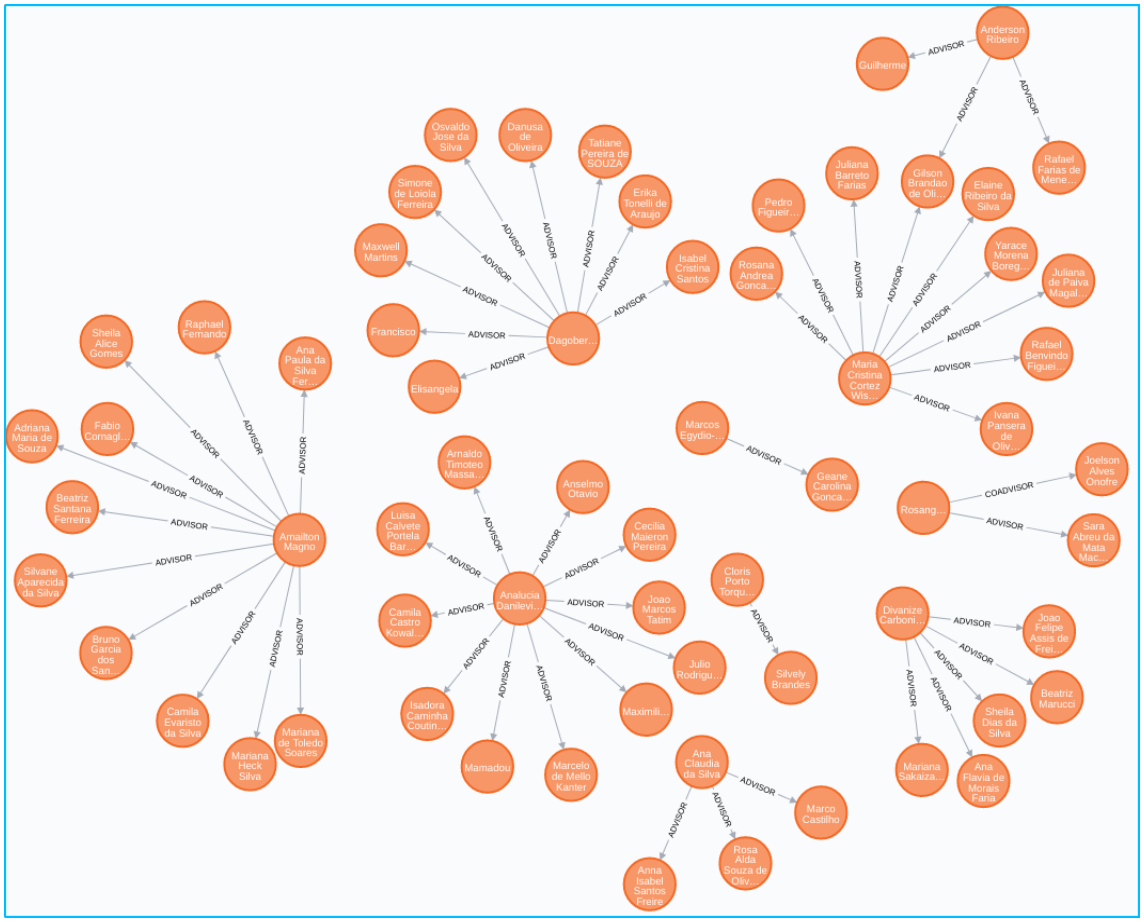

Figure 4. Student–(co-)advisor graph drawing.

We have described how programming and many other computer science concepts like regular expressions, graph modeling, domain-specific query languages and information visualizations were important for extracting and combining data, producing new pieces of information that can be appropriately presented. In terms of methodology, dividing the work in preprocessing and question-answering tasks allows to prepare all necessary data for supporting the more semantic and thoughtful design activities. It is worth noting, though, that the development process is cyclic: the desirable final visualizations demand data preprocessing tasks, while limitations in these initial stages impose constraints on what can be obtained at the end. Thus, some preprocessing demands were only sketchily approached here. For instance, name disambiguation was done in a simple way using Levenshtein similarity measure. However, this is not enough for solving many cases, since the entrance of people names in the BDTD database occurs by distinct employees and even institutions. Better approaches for dealing with name disambiguation in the BDTD are necessary, and we are working on this as a future work. Another problem that could not be tackled completely is the low interactivity of some visualizations built in the scope of this project. The Jupyter Notebook used for creating, presenting and discussing two of the proposed solutions is very didactic. But, it lacks the interactivity that some of the above mentioned tools have for the manipulation of information visualizations. The visualizations *per si* do not answer the research questions. They are just a means by which the researcher's cognitive system can be expanded to detect patterns,

allowing him or her to see information that was previously hidden. Interactions for filtering data, choosing more effective visual mappings and looking at the data from different perspectives are essential for the success of that goal. New investigations should consider these aspects.

A very interesting idea explored in the present study was the use of a knowledge graph representation of BDTD data, supported by the Neo4j tool. This was implemented at a more advanced stage of work development, but it showed to be promising for a much broader context. For example, all preprocessing tasks could feed the knowledge graph directly, instead of generating modified CSVs as an intermediate step. Moreover, it is possible to include in the knowledge graph all the data available in the original CVSs, making it a self-sufficient source for creating any visualization. Complex network analysis algorithms can even be applied to improve the information in the knowledge graph.

Finally, the study has resulted in visualizations that are now been used by the social sciences researches for supporting the necessary insights on the Brazilian academic production in the investigated themes. The work also was useful for identifying problems in the BDTD database, such as: repeated dissertations/thesis entrance within the same theme but with different identification codes or year of publication; absence of data in some cells; lack of a regular format for several fields; repetition or swap of information between columns, character encoding losses; and, eventually, columns without the proper character delimiter. Despite such obstacles and the heterogeneity of the database, up to now it has been possible to process and to treat most of the incorrect information.

References

Alves, A. D., Yanasse, H. H., and Soma, N. Y. (2011). SUCUPIRA: A system for Information extraction of the Lattes Platform to identify academic social networks. In *6th Iberian Conference on Information Systems and Technologies (CISTI 2011)*, pages 1–6. ISSN: 2166-0735.

Bucher-Maluschke, J. S. N. F., Silva, J. C. e., and Souza, I. B. d. S. d. (2019). REVISÃO SOBRE O PRESÍDIO FEMININO NOS ESTUDOS BRASILEIROS. *Psicologia & Sociedade*, 31. Publisher: Associação Brasileira de Psicologia Social.

Campello, B. S., Vianna, M. M., Caldeira, P. d. T., Abreu, V. L. F. G., Carvalho, M. d. C., and Benigno, A. C. e. S. (2007). Literatura sobre biblioteca escolar: características de citações de teses e dissertações brasileiras. *Transinformação*, 19(3):227–236.

Costa, O. A. d. and Rodrigues, A. C. L. (2019). Mapeamento da produção científica na BDTD do IBICT sobre a Pedagogia da Alternância de 2011 a 2018. *Revista Brasileira de Educação do Campo*, 4:e7257–e7257.

Hayashi, M. C. P. I., Cabrero, R. d. C., Costa, M. d. P. R. d., and Hayashi, C. R. M. (2007). Indicadores da participação feminina em Ciência e Tecnologia. *Transinformação*, 19(2):169–187. Publisher: Pontifícia Universidade Católica de Campinas.

Kambatla, K., Kollias, G., Kumar, V., and Grama, A. (2014). Trends in big data analytics. *Journal of Parallel and Distributed Computing*, 74(7):2561–2573.

McAfee, A. and Brynjolfsson, E. (2012). Big data: the management revolution. *Harvard Business Review*, 90(10):60–66, 68, 128.

Mena-Chalco, J. P. and Cesar Junior, R. M. (2009). ScriptLattes: an open-source knowledge extraction system from the Lattes platform. *Journal of the Brazilian Computer Society*, 15(4):31–39.

Santos Junior, J. d. S. and Real, G. C. M. (2017). Dropout from higher education: the state of the art of researches in Brazil since 1990. *Avaliação: Revista da Avaliação da Educação Superior (Campinas)*, 22(2):385–402. Publisher: Avaliação: Revista da Avaliação da Educação Superior.

Silva, P. d., Pinto, G. F. S., and Furnival, A. C. (2018). Análise dos aspectos normativos e legais do uso de fotografias contidas em teses e dissertações disponíveis na BDTD/IBICT. *Brazilian Journal of Information Science: Research Trends*, 12(3):22 ao 33–22 ao 33. Number: 3.

Vitta, F. C. F. d., Sgavioli, A. J. R., Scarlassara, B. S., Novaes, C. F. M., Cruz, G. d. A., and Moura, M. M. (2018). National Scientific Production in the Special Education Area and Daycare. *Revista Brasileira de Educação Especial*, 24(4):619–636. Publisher: Associação Brasileira de Pesquisadores em Educação Especial.

Zhang, J., Wong, J.-S., Li, T., and Pan, Y. (2014). A comparison of parallel large-scale knowledge acquisition using rough set theory on different MapReduce runtime systems. *International Journal of Approximate Reasoning*, 55(3):896–907.