The Gold Tree: An Information System for Analyzing Academic Genealogy

Gabriel Madeira, Eduardo N. Borges, Matheus Barañano, Prícilla Karen Nascimento,
Giancarlo Lucca, Maria de Fatima Maia, Helida Salles and Graçaliz Dimuro
Centro de Ciências Computacionais, Universidade Federal do Rio Grande – FURG,

Keywords: Academic Genealogy, Genealogy Trees, Data Integration, Information Visualization.

Abstract: Academic genealogy investigates the relationships between student researchers and academy professionals. In recent years, it proved to be a powerful technique to help analyze the spread of scientific knowledge. Tools that make to visualize these relationships among academics easier are potentially useful and have been proposed. This work specifies and describes the development of a Web information system for creating and visualizing academic genealogy trees from a set of metadata extracted and integrated from multiple sources. The proposed system allows a researcher to query and track information about his or her advisers and graduate students at any level. A case study was explored to validate the system using data from more than 570 thousand theses and dissertations.

1 INTRODUCTION

Currently, there are a large number of scientific publications and academic papers available in various Web repositories. Each research institution or university publishes the results achieved in its own institutional repository. In this way, scientific publications are cataloged and organized in a dispersed way. These data together contain the major scientific contributions and collaborations among researchers over time. Analyzing the metadata from multiple publications allows mapping and understanding how the relationships between researchers affect the advancement of knowledge in several areas of science.

Genealogy is an auxiliary field of history that studies the origin, evolution and spread of family groups (Malmgren RD, 2010). This evolution is often represented using a structured diagram in the form of family trees (David and Hayden, 2012). Genealogy trees are well-known structures that organize, through kinship ties, the whole history of an individual's ancestors. Using this structure we can analyze the origin and development over time of the lineage of a family.

Genealogy trees can be used in academia to analyze relationships between professors, students, and researchers. Figure 1 shows an example of academic genealogy tree. Any metadata sets that describe the elements or their relationships can be used.

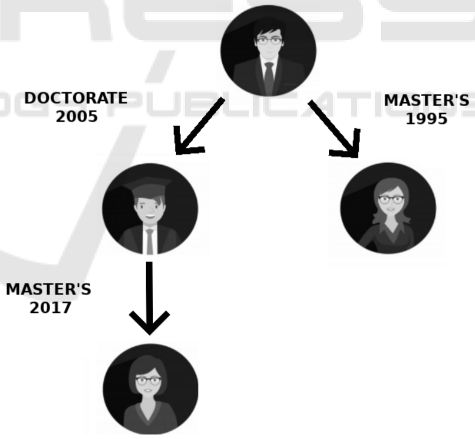


Figure 1: An academic genealogy tree representing advising relationships in graduate programs. Each edge starts from the advisor and presents the year of master's or doctoral degree.

When drawing an academic genealogy tree, it is possible to see who guided a researcher and how he or she influenced other researchers over time. Getting a forest (set of trees) makes possible to describe a research area using metrics that allow, through statistical analyzes and data mining, to extract relevant knowledge for the area under study (Chang, 2011).

Madeira, G., Borges, E., Barañano, M., Nascimento, P., Lucca, G., Maia, M., Salles, H. and Dimuro, G.
The Gold Tree: An Information System for Analyzing Academic Genealogy.
DOI: 10.5220/0007758401140120
In *Proceedings of the 21st International Conference on Enterprise Information Systems (ICEIS 2019)*, pages 114-120
ISBN: 978-989-758-372-8

Therefore, these structures allow us to analyze how knowledge is spreading across generations of scientists and how these links affect the development of science.

This paper describes an information system called The Gold Tree, which goal is to visualize academic genealogy trees created from a set of metadata extracted and integrated from multiple sources. The Information Management Research Group developed it in Centro de Ciências Computacionais at Universidade Federal do Rio Grande (FURG). The proposed system allows a researcher to query and track information about his or her advisers and graduate students at any level. A case study was explored to validate the system using data from more than 570 thousand theses and dissertations.

The rest of this paper is organized as follows. In Section 2, we discuss related work. In Section 3, we present the methodology to develop the proposed solution. In Section 4, we give details on the obtained results. Finally, in Section 5, we draw our conclusions and point out some future work directions.

2 RELATED WORK

In recent years, several studies have explored the visualization of academic collaboration data. While some platforms such as ResearchGate (Yu et al., 2016), Google Citations, and the Web of Science (WoS) classify registered researchers by citation indexing for their articles and papers (Barabâsi et al., 2002), other tools such as Pajek (Batagelj and Mrvar, 2002) and PubNet (Douglas et al., 2005) are only concerned with viewing research networks. Furthermore, we point out that there are also solutions that use specific data sources to extract information and generate knowledge from co-authoring relationships (Mena-Chalco and Cesar-Jr, 2013; Laender et al., 2011). The following subsections present in detail the work used as baseline in the validation of the proposed system.

2.1 Academic Family Tree

Neurotree is a Web database created to document the lineage of academic mentorship in neuroscience (David and Hayden, 2012). The authors present a temporal analysis of the database growth in a period of seven years. The following metric were performed: the number of researchers and relationships, the monthly growth rate, the fraction of researchers linked in the main graph, the average distance between researchers, and the average number of connections per researcher. In addition, they report the accuracy of related data in Neurotree with data reported on Web sites of five research groups. Finally, in order to study the relationship between mentorship groups and research areas within neuroscience, they provide a clustering analysis.

This tree exists as a part of the larger Academic Family Tree[1], which seeks to build a genealogy across multiple academic fields, building a single, interdisciplinary academic genealogy. Figure 2 present the result of a query by research name.

The contents of the database are entirely crowdsourced. So it is totally dependent on human effort. This feature makes it very susceptible to field fill errors, as well as always presenting incomplete data. Any Web user can add information about researchers and the connections between them, which can leave the database with poor quality and with false information.

2.2 Acácia Plataform

The Acácia Platform[2] (Damaceno, 2017) is a system created in 2017 for documenting the formal relations of advising in the context of the Brazilian graduate programs. The system uses data registered in the Lattes Platform[3], which is a database of Brazilian researchers' curricula maintained by the Ministry of Science and Technology and Innovation. Currently, the Acacia Platform has over 1 million vertices and relationships. Each vertex represents a researcher and each edge an advising relation completed between two researchers (advisor and student). Figure 3 present the result of a query by research name. The system shows some bibliometric indexes as the number of direct and indirect descendants and information about the advising relationships.

2.3 Science Tree

Created in 2015, the Science Tree [4] application collect metadata of academic genealogy from many countries (Dores et al., 2016). The authors are crawling data from a variety of sources, including the Networked Digital Library of Theses and Dissertations (NDLTD), which has more than 4.5 million theses and dissertations from around the world. They develop a framework to extract academic genealogy trees from this data and, providing a series of analyses that describe the main properties of the academic genealogy

[1] https://academictree.org

[2] http://plataforma-acacia.org

[3] http://lattes.cnpq.br

[4] http://www.sciencetree.net

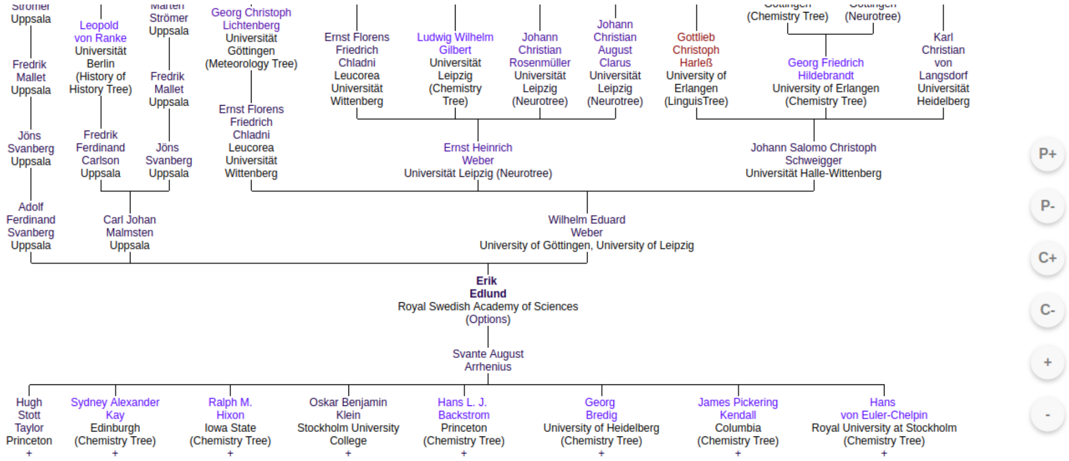
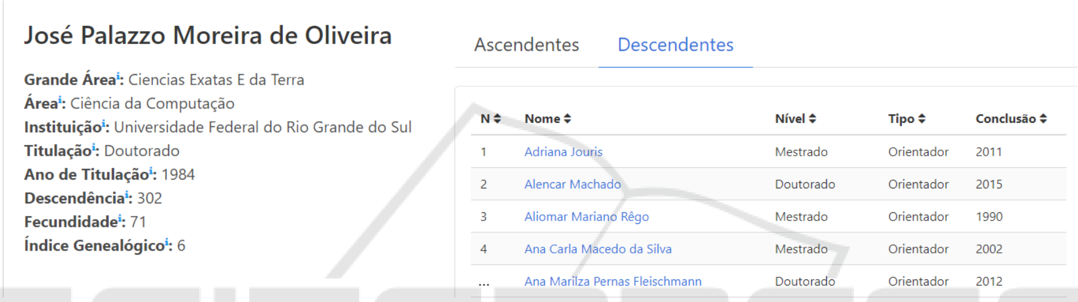

Strömer Uppsala
Leopold von Ranke Universität Berlin (History of History Tree)
Mårten Strömer Uppsala
Georg Christoph Lichtenberg Universität Göttingen (Meteorology Tree)
Ernst Florens Friedrich Chladni Leucorea Universität Wittenberg
Ludwig Wilhelm Gilbert Universität Leipzig (Chemistry Tree)
Johann Christian Rosenmüller Universität Leipzig (Neurotree)
Johann Christian August Clarus Universität Leipzig (Neurotree)
Gottlieb Christoph Harleß University of Erlangen (LinguisTree)
Göttingen (Chemistry Tree)
Göttingen (Neurotree)
Georg Friedrich Hildebrandt University of Erlangen (Chemistry Tree)
Karl Christian von Langsdorf Universität Heidelberg
Fredrik Mallet Uppsala
Fredrik Mallet Uppsala
Ernst Florens Friedrich Chladni Leucorea Universität Wittenberg
Jöns Svanberg Uppsala
Fredrik Ferdinand Carlson Uppsala
Jöns Svanberg Uppsala
Ernst Heinrich Weber Universität Leipzig (Neurotree)
Johann Salomo Christoph Schweigger Universität Halle-Wittenberg
Adolf Ferdinand Svanberg Uppsala
Carl Johan Malmsten Uppsala
Wilhelm Eduard Weber University of Göttingen, University of Leipzig
Erik Edlund Royal Swedish Academy of Sciences (Options)
Svante August Arrhenius
Hugh Stott Taylor Princeton
Sydney Alexander Kay Edinburgh (Chemistry Tree)
Ralph M. Hixon Iowa State (Chemistry Tree)
Oskar Benjamin Klein Stockholm University College
Hans L. J. Backstrom Princeton (Chemistry Tree)
Georg Bredig University of Heidelberg (Chemistry Tree)
James Pickering Kendall Columbia (Chemistry Tree)
Hans von Euler-Chelpin Royal University at Stockholm (Chemistry Tree)
P+
P-
C+
C-
+
-

Figure 2: Result of the query “Erik Edlund” using the Academic Family Tree.



José Palazzo Moreira de Oliveira

Grande Área: Ciencias Exatas E da Terra
Área: Ciência da Computação
Instituição: Universidade Federal do Rio Grande do Sul
Titulação: Doutorado
Ano de Titulação: 1984
Descendência: 302
Fecundidade: 71
Índice Genealógico: 6

Ascendentes Descendentes

Nº	Nome	Nível	Tipo	Conclusão
1	Adriana Jouris	Mestrado	Orientador	2011
2	Alencar Machado	Doutorado	Orientador	2015
3	Aliomar Mariano Rêgo	Mestrado	Orientador	1990
4	Ana Carla Macedo da Silva	Mestrado	Orientador	2002
...	Ana Marilza Pernas Fleischmann	Doutorado	Orientador	2012

Figure 3: Result of the query “José Moreira Oliveira Palazzo” using the Acácia Platform. For each relationship the name, academic degree and year of conclusion are presented.

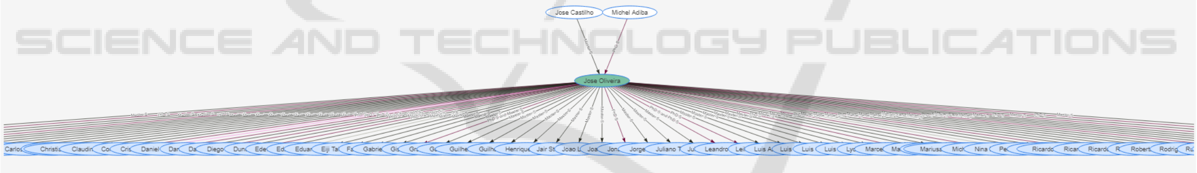


Figure 4: Result of the query “José Moreira Oliveira Palazzo” using the Science Tree.

tree. Figure 4 present the result of a query using the same researcher of Figure 3.

3 METHODOLOGY

This section presents the methodology adopted in this article. The proposed approach is divided into the following steps: data source definition, harvesting and pre-processing, data modeling and indexing, and web information system construction.

Unlike the Academic Family Tree related work presented in the 2.1 section, which is user dependent, we have chosen to collect official data available in digital libraries. These data sources must support some interoperability feature such as the OAI-PMH protocol (Lagoze and Van de Sompel, 2001) and the Dublin Core[5] format. This choice solves the problem of cold start and registration of false information.

In order to evaluate the proposed system we use the Brazilian Digital Library of Theses and Dissertations[6] (BDTD). Developed and managed by the Brazilian Institute of Information in Science and Technology (IBICT), it integrates the repositories of educational and research institutions in Brazil, and also stimulates the registration and publication of theses and dissertations electronically. At the time of our harvesting more than 570,000 documents were indexed.

The main metadata fields collected were:

- author;

[5] http://dublincore.org
[6] http://bdtd.ibict.br

- advisor;
- name of the educational or research institution;
- acronym of the educational or research institution;
- title;
- topics;
- URL of the document in the original repository;
- PhD thesis or Master's dissertation;
- URL of the author curriculum in Lattes Plataform;
- citation;
- year of publication.

After the data selection and harvesting, a set of cleaning operations were applied. Several analyzes were performed to identify anomaly patterns, so errors were corrected or eliminated. Some transformation operations were applied on the author and advisor fields. The most common ones include inverting last names based on the comma character, removing institution acronyms, and removing structural prefixes present in the content of a metadata. In addition, duplicate tuples from more than one digital repository were removed. At the end of the cleaning process, the number of theses and dissertations decreased to 465,847.

In Computer Science, trees are data structures widely used to represent elements hierarchically organized. However, the academic genealogy trees are represented by graphs, because a researcher may have more than one ascending (one for each dissertation or thesis) and because there may be cycles. That's why we transformed the cleaned data to generate two tables in the relational model. The first table contains all the academics and their properties. The second keeps M:N directed advising relationships between pairs of academics. Figure 5 shows this data model. The Database Management System (DBMS) used for storage was PostgreSQL. All the process of harvesting and pre-processing was developed in the programming language PHP.

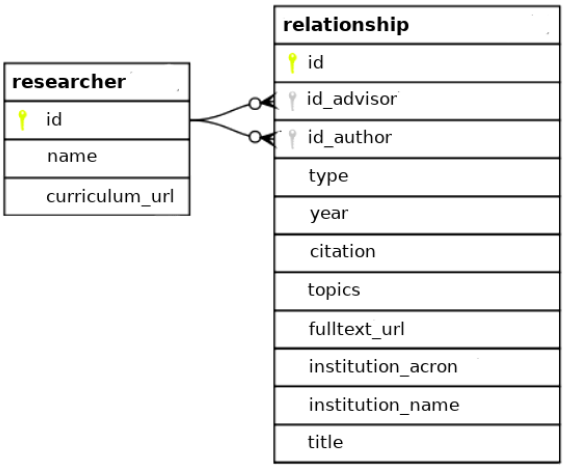


Figure 5: Relational data model representing the genealogy academic trees.

The Web application interface was developed using HTML, CSS and Javascript. We have used the dagre-d3 library (Roeder, 2018) to draw the academic trees. The queries are sent to the back-end using Jquery library (Osmani, 2012). Methods implemented in PHP retrieve the information stored in the DBMS. Figure 6 shows the data flow between front and back-end.

4 RESULTS

Figure 7 shows the web interface of the proposed system, which is designed to be simple and intuitive. The button *Search Academic* opens a query field that allows the user to search by the name of the researcher.

The architecture presented in Figure 6 allowed to implement a dynamic search that suggests multiple researchers as the user types in the query field. For each character entered, the results are filtered and displayed on the screen. All substrings of author and advisor names with at least 3 characters has been indexed in the DBMS, so the user does not need to know the full name of a researcher, nor the order of names or to complete each name. Figure 8 exemplifies this behavior while the user are querying by "avancini mar".

Next to the button *Search Academic* the user can edit the depth level. The tree expands that number of levels both toward the leaves, and toward the root. Thus, he or she can see the advising lineage and the graduated students. Figure 7 shows the result tree selecting "Rita Maria Pereira Avancini" among the five returned researches and setting two levels.

Each researcher is represented by a vertex that contains his or her full name. Master's advising relationships are edges with M-YEAR labels in blue. Doctoral advising have D-YEAR labels in red.

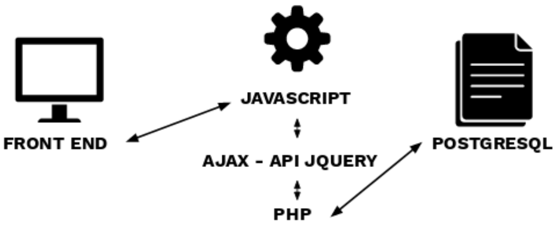


Figure 6: Data flow between front and back ends.

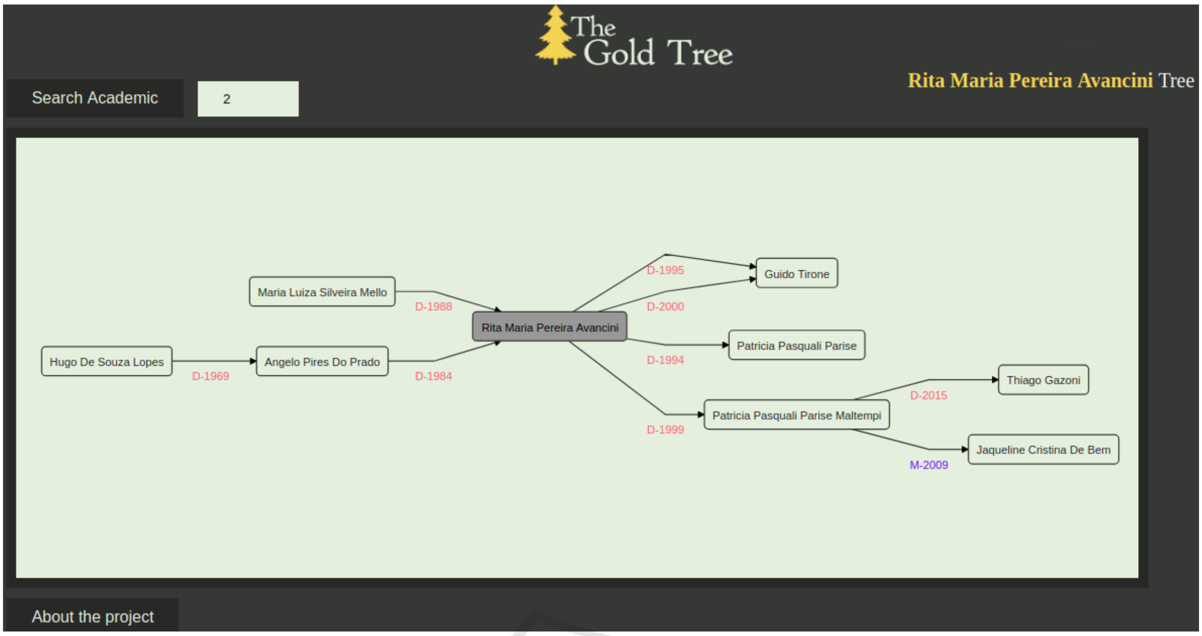
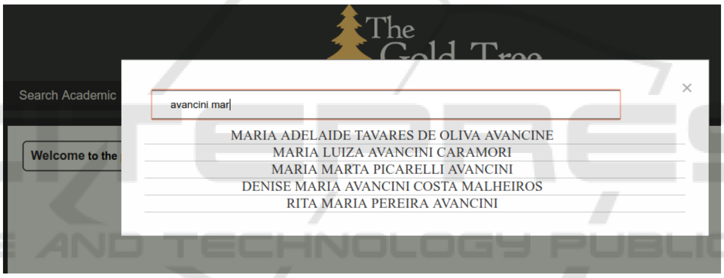


Figure 7: Web interface of the proposed system showing the result selecting “Rita Maria Pereira Avancini” with 2 levels.



avancini mar

MARIA ADELAIDE TAVARES DE OLIVA AVANCINE
MARIA LUIZA AVANCINI CARAMORI
MARIA MARTA PICARELLI AVANCINI
DENISE MARIA AVANCINI COSTA MALHEIROS
RITA MARIA PEREIRA AVANCINI

Figure 8: Example of the dynamic search feature, that suggests multiple researchers as the user types in the query field.

The user can freely move the tree in the rectangular area in which it appears and zoom in/out using the mouse controls. By clicking on a relationship, a window opens displaying the information available in the thesis or dissertation metadata (Figure 9). Also, when you click on a vertex, a new tree is generated using the selected researcher as the target of the query.

The information system developed is available online[7].

5 CONCLUSION

This work presents a study on tools for visualization of academic genealogy, besides proposing the development of an unprecedented and effective system.

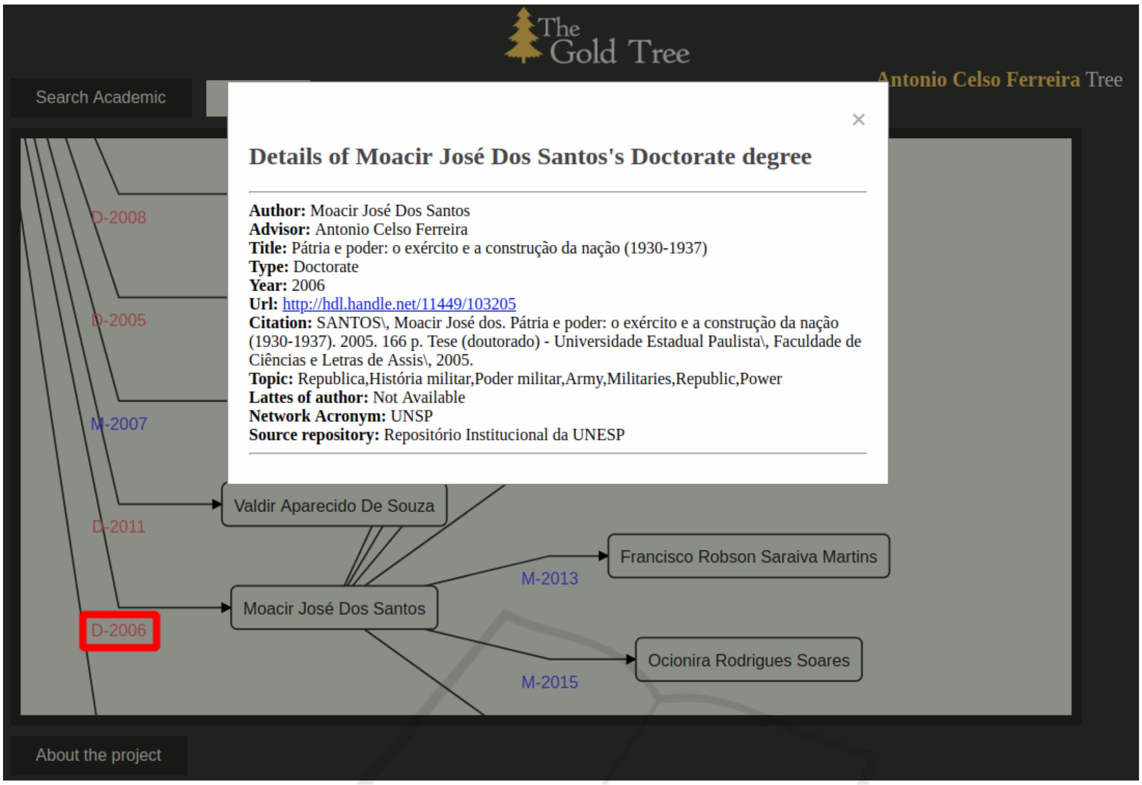
The Gold Tree is based on official information stored in the Brazilian Digital Library of Theses and Dissertations. However, the data source chosen can easily be replaced by one that publishes the metadata using the Dublin Core standard and the OAI-PHM protocol.

The information system described in this paper can assist in research that investigates relationships of academic genealogy to reach the most diverse objectives. With the user-friendly search interface, users can analyze and draw conclusions from multiple tree views.

Compared with related work presented in Section 2, it is clear that the proposed system has several advantages. We use a reliable data source to map advising, unlike Academic Family Tree where these relationships are entered by users. In Acácia Platform, although data is extracted from Lattes Platform, it is not possible to visualize the genealogy trees. Finally, when compared to Science Tree, our system allows to build trees with any depth levels, in addition to allowing the dynamic search by the researcher name.

Future work includes the implementation of a routine for updating the collected data, the incorporation

[7] http://thegoldtree.c3.furg.br/



Details of Moacir José Dos Santos's Doctorate degree

Author: Moacir José Dos Santos
Advisor: Antonio Celso Ferreira
Title: Pátria e poder: o exército e a construção da nação (1930-1937)
Type: Doctorate
Year: 2006
Url: http://hdl.handle.net/11449/103205
Citation: SANTOS\, Moacir José dos. Pátria e poder: o exército e a construção da nação (1930-1937). 2005. 166 p. Tese (doutorado) - Universidade Estadual Paulista\, Faculdade de Ciências e Letras de Assis\, 2005.
Topic: Republica,História militar,Poder militar,Army,Militaries,Republic,Power
Lattes of author: Not Available
Network Acronym: UNSP
Source repository: Repositório Institucional da UNESP

Figure 9: Metadata describing the thesis or dissertation presented when clicking on an advising relationship.

of other international data sources, improvements of interface, the inclusion of bibliometric indexes and the implementation of the search for other metadata besides the name of the academic.

In addition, we intend to implement an expert system that aims to suggest doctoral advising relationships based on the advisors' research subjects. The system will be able to extract several features from the titles and abstracts of the theses and dissertations of all the descendants of a researcher to set up the advising profile. Using multiple machine learning algorithms, a learned profile will be assigned to a doctoral candidate with certain confidence to make the choice of an advisor easier. Combining multiple classifiers with high diversity will improve the quality of the system.

ACKNOWLEDGEMENTS

This research is partially supported by CNPq (Proc. 307781/2016-0 & 305882/2016-3) and FAPERGS (TO 17/2551-0000872-3). Special thanks to Washington Luís R. de Carvalho Segundo, IBICT employee who helped metadata harvesting.

REFERENCES

Barabâsi, A.-L., Jeong, H., Néda, Z., Ravasz, E., Schubert, A., and Vicsek, T. (2002). Evolution of the social network of scientific collaborations. *Physica A: Statistical mechanics and its applications*, 311(3-4):590–614.

Batagelj, V. and Mrvar, A. (2002). Pajek – analysis and visualization of large networks. In *Graph Drawing*, pages 477–478, Berlin, Heidelberg. Springer.

Chang, S. (2011). *Academic genealogy of mathematicians*. World Scientific.

Damaceno, R. J. P.; Rossi, L. M.-C. J. P. (2017). Identificação do grafo de genealogia acadêmica de pesquisadores: Uma abordagem baseada na plataforma lattes. In *Proceedings of the 32nd Brazilian Symposium on Databases, Uberlândia*.

David, S. V. and Hayden, B. Y. (2012). Neurotree: A collaborative, graphical database of the academic genealogy of neuroscience. *PLoS ONE*, 7(10):e46608+.

Dores, W., Benevenuto, F., and Laender, A. H. (2016). Extracting academic genealogy trees from the networked digital library of theses and dissertations. In *Proceedings of the 16th ACM/IEEE-CS on Joint Conference on Digital Libraries*, pages 163–166, New York, NY, USA. ACM.

Douglas, S. M., Montelione, G. T., and Gerstein, M. (2005). Pubnet: a flexible system for visualizing literature derived networks. *Genome Biology*, 6(9).

Laender, A., Moro, M., Silva, A., et al. (2011). Ciência brasil-the brazilian portal of science and technology. In *Seminário Integrado de Software e Hardware*, pages 1366–1379. Sociedade Brasileira de Computação.

Lagoze, C. and Van de Sompel, H. (2001). The open archives initiative: Building a low-barrier interoperability framework. In *Proceedings of the 1st ACM/IEEE-CS Joint Conference on Digital ibraries*, pages 54–62, New York, NY, USA. ACM.

Malmgren RD, Ottino JM, N. A. L. (2010). The role of mentorship in protégé performance. *Nature Internacional Journal of Science*, 465.

Mena-Chalco, J. P. and Cesar-Jr, R. M. (2013). *Prospecção de dados acadêmicos de currículos Lattes através de scriptLattes*, chapter Bibliometria e Cientometria: reflexões teóricas e interfaces, pages 109–128. Pedro & João Editores, São Carlos.

Osmani, A. (2012). *Learning JavaScript Design Patterns: A JavaScript and jQuery Developer's Guide*. O'Reilly Media, Inc.

Roeder, L. (2018). Dagre-d3. https://github.com/dagrejs/dagre-d3.

Yu, M.-C., Wu, Y.-C. J., Alhalabi, W., Kao, H.-Y., and Wu, W.-H. (2016). Researchgate: An effective altmetric indicator for active researchers? *Computers in Human Behavior*, 55:1001–1006.