CiênciaBrasil - The Brazilian Portal of Science and Technology

Alberto H. F. Laender[1], Mirella M. Moro[1], Altigran S. Silva[2], Clodoveu A. Davis Jr.[1], Marcos André Gonçalves[1], Renata Galante[3], Allan J. C. Silva[1], Carolina A. S. Bigonha[1], Daniel Hasan Dalip[1], Eduardo M. Barbosa[1], Eduardo N. Borges[3], Eli Cortez[2], Peterson Procópio Jr.[1], Rafael Odon de Alencar[1], Thiago N. C. Cardoso[1], Thiago Salles[1]

[1]Departamento de Ciência da Computação
Universidade Federal de Minas Gerais (UFMG), Belo Horizonte, MG

[2]Departamento de Ciência da Computação
Universidade Federal do Amazonas (UFAM), Manaus, AM

[3]Instituto de Informática
Universidade Federal do Rio Grande do Sul (UFRGS), Porto Alegre, RS

{laender,mirella,clodoveu,mgoncalv,allan,carolb,hasan,
emb,peterson,odon,thiagon,tsalles}@dcc.ufmg.br,
{alti,eccv}@dcc.ufam.edu.br, {galante,enborges}@inf.ufrgs.br

Abstract. *Research social networks are a potentially useful resource for studying science and technology indicators from specific communities (e.g., a country). However, building and analyzing such networks beget challenges beyond those from regular social networks, since data about actors and their relationships are usually dispersed across various sources. In this paper, we present a research social network built from an individual perspective by gathering data from a Brazilian curricula vitae repository. We describe its architecture and the solutions adopted for data collection, extraction and deduplication, and for materializing and visualizing the network.*

1. Introduction

In July 2008, CNPq (the Brazilian National Council for Scientific and Technological Development) lauched a research program called National Institutes of Science and Technology (or simply INCT)[1]. The INCT program has become a powerful initiative for promoting science, technology and innovation in the country. Its main goal is to mobilize and aggregate in networks the best research groups in frontier areas, which cover a wide range of topics, including agribusiness, health, energy, nanotechnology, computing and social sciences. Among those institutes, the INWeb[2] (INCT for the Web) focuses on cutting edge research, technology transfer, education of highly qualified human resources and dissemination of science and technology on Web-related topics. One of its projects, *CiênciaBrasil* – the *Brazilian Portal of Science and Technology*[3] – consists of building and analyzing a *Research Social Network* that allows studying science and technology indicators (knowledge production, research collaboration, human resources formation, technology transfer etc.) involving Brazilian researchers within the INCT program.

[1]INCT Program http://www.cnpq.br/programas/inct/_apresentacao
[2]InWeb http://www.inweb.org.br
[3]CiênciaBrasil http://pbct.inweb.org.br

Gathering and processing information for building and analyzing research networks beget challenges that go beyond those of regular social networks, such as Facebook and MySpace[4] (see [Tang et.al 2008]). The data is usually not provided by the network members themselves (i.e., researchers) but must be collected and extracted from various sources (including the Web). Moreover, the existing relationships among these members are generally implicitly embedded in the data and must be derived by processing them (e.g., using sophisticated automatic extraction techniques). Even if the data can be correctly obtained, data duplication and (name) ambiguity may occur since such data is usually collected from several heterogeneous sources. In order to process such a network with a minimum of quality, we need to deal with those issues as well as others associated with the visualization, navigation and analysis of specific instances of the network. Examples of Web applications that exploit academic networks and face those challenges are ArnetMiner[5] and Microsoft Academic Search[6].

In Brazil, *Lattes*[7] is a Web platform made available by CNPq for storing, managing and searching for curricula vitae of researchers involved with Brazilian institutions. With Lattes, researchers (from undergrads to seniors) inform their academic achievements including education, professional history, publications, funding, academic positions, advising, awards, etc. Lattes has been recognized as one of "the cleanest researcher databases in existence" [Lane 2010]. With such organized data, Lattes became the natural source for building *CiênciaBrasil.* Indeed, the existence of a standardized platform for curricula vitae publication provides an opportunity for building a large, *individually centered* repository of scientific and educational information. From the individual perspective, it is possible to put together research groups and associations, as well as to build a fairly reliable research social network covering several knowledge fields.

However, obtaining and organizing data from Lattes is a challenge per se. When using Lattes, researchers have to fill out a form with predefined fields to include information about their scientific production. This is done in an isolated way and, when published as a Web page, there are no clear delimiters to (for example) easily extracting the desired fields from publication citations (e.g., coauthor names, title, venue). Hence, it requires sophisticated extraction strategies. Moreover, since several coauthors include the same publications in their curricula vitae, group and network statistics require a *deduplication* step in order to generate precise production figures. Deduplication is known to be a difficult problem [Carvalho et al. 2008, Geer 2008, Sarawagi and Bhamidipaty 2002], which happens not only due to possible errors in filling out forms, but also due to non-standardized inclusion of coauthor information.

In this paper we introduce *CiênciaBrasil* and provide an overview of our solutions for some of the aforementioned challenges. We also illustrate the potential of *CiênciaBrasil* as a platform for large scale academic research evaluation and analysis. Furthermore, all challenges presented in this work are related to two of SBC's Grand Challenges in Computer Science Research in Brazil[8]. Specifically, it is related to the first

[4]Facebook http://www.facebook.com; MySpace http://www.myspace.com
[5]ArnetMiner http://www.arnetminer.com
[6]Microsoft Academic Search http://academic.research.microsoft.com
[7]Lattes http://lattes.cnpq.br/english
[8]SBC http://www.sbc.org.br

challenge ("Management of information over massive volumes of distributed multimedia data") due to the big volume of data extracted and evaluated for building the networks and to the fourth challenge ("Participative and universal access to knowledge for the Brazilian citizen") since the visualization and analysis of the research networks will allow to identify and disseminate information that is related to the Brazilian research, bringing it closer to the regular (i.e., non-scientist) citizen.

The contributions of this paper can be then summarized as follows. After discussing related work (Section 2), we introduce an architecture for building academic networks from an individual perspective. In a nutshell, with such an architecture, we collect and treat data from individuals and then build the network. Also, we use specific algorithms for collecting and extracting data as well as deduplicating it. Then, we build the research network and provide different ways to visualize it (Section 3). We also present the major features of *CiênciaBrasil* and discuss examples of different visualization resources provided on its portal (Section 4).

2. Related Work

This paper presents *CiênciaBrasil*, a Web portal for analyzing and viewing information on science and technology. Specifically, *CiênciaBrasil* gathers data from Brazilian researchers, integrates them and generates distinct individual and group-based visualizations that can aid on different analysis and evaluation processes. In order to generate such views, *CiênciaBrasil* employs social networks techniques that are tailored for the research context. This section discusses some tools that are similar or that employ concepts from social networks.

ArnetMiner [Tang et.al 2008] provides online search and mining services for researcher social networks. It creates profiles for researchers based on Web data that includes bibliographic data and researcher's data from multiple sources. With the networks, it also allows to discover patterns, to find experts, conferences and papers on specific topics, and to rank people based on their research achievements.

Microsoft Academic Search is based on the concept of object-level vertical search [Nie et al. 2005] and provides different ways to explore scientific papers, conferences, journals, and authors. Specifically for authors (the focus of our work), it shows a profile page with affiliation, publications, citations, g-index, h-index, coauthors, and coauthorship graph. The data is acquired from Web sources including DBLP, ACM Digital Library, CiteSeer, among others.

Both ArnetMiner and MS Academic Search extract researchers' profiles by integrating publication information, for example from digital libraries. Then, they build academic networks and provide search services. However, extracting such data from the Web and integrating them can be error prone: ArnetMiner allows registered users to correct errors through an editing interface and Microsoft Academic Research has serious problems with name disambiguation. For example, our coauthor Clodoveu Davis Jr. has *five* entries for different profiles (in March 2011).

SciVerse Scopus[9] is one of the largest abstract and citation data source available for research literature. Among its features, it has an author search service that provides

[9]Scopus http://www.scopus.com

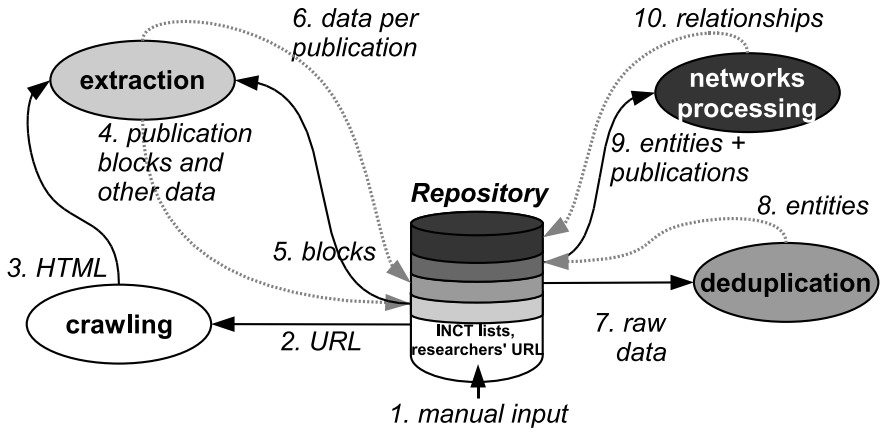


Figure 1. *CiênciaBrasil* Architecture

a full profile of authors, which includes name, affiliation, references, citations, h-index, coauthors, subject area, and history. SciVerse Scopus is similar to *CiênciaBrasil* in the sense that it builds author profiles based on its repository. However, its repository is actually a collection of external sources, and it incurs on problems such as disambiguation (two different people with same name) and deduplication (the same person with two different names). For example, our coauthor Alberto H. F. Laender has *three* entries, even though it correctly found three forms of writing his name (Alberto H. F. Laender, A. H. F. Laender and Alberto Laender). Nonetheless, SciVerse Scopus also offers other services such as the Journal Analytics that are beyond *CiênciaBrasil* scope.

3. *CiênciaBrasil* Architecture

This section introduces the general architecture employed by *CiênciaBrasil* to build a research social network from an individual perspective. In a nutshell, with this architecture, information from individuals is collected and treated to, then, build the network based on it. Specifically, the whole process starts with the manual data input of the INCTs basic data (Arrow 1 in Figure 1). The basic data is composed of URLs to the Lattes curricula vitae of the researchers. From the URLs, the curricula vitae pages are collected (Arrows 2 and 3). The process follows by extracting and organizing the necessary information (Arrows 4 to 6), which then is deduplicated (Arrows 7 and 8) and disambiguated (Arrow 9) before building the network (Arrow 10). Next, we present the technical and scientific details of each process within the architecture.

3.1. Crawling

The first process is to manually input the basic data from each INCT into the repository. The data include the name, acronym, principal investigators, home institution, other collaborating institutions, etc, and the list of Lattes URLs of its participating researchers (Arrow 1 in Figure 1). The first version of *CiênciaBrasil* covers 4,676 researchers from 123 INCTs. The input list of URLs is then fed to the crawling process (Arrow 2 in Figure 1), for collecting the respective curricula vitae from the Lattes platform.

The crawling process applies an asynchronous concurrent download strategy to increase throughput (e.g., due to the high latency observed when establishing/negotiating HTTP connections, handling several connections simultaneously ultimately maximizes

bandwidth usage and increases the number of crawled curricula vitae's per time unit). It also keeps track of download errors to avoid any possible data loss [Heydon and Najork 1999]. After crawling, the gathered data is available in pure HTML format (Arrow 3 in Figure 1), which will be further decomposed in finer-grained structures called *information blocks*.

3.2. Data Extraction

The data extraction process involves two steps. First, it extracts *information blocks* from the HTML pages, i.e., blocks containing relevant information for building the research social network. Then, it extracts specific finer-grain data from the information blocks, such as publication metadata (e.g., author names, title, venue, etc.).

In the first step, despite the standardized structure of the Lattes curriculum vitae (with predefined fields filled by the researchers) there are no explicit delimiters to guide the extraction process. For example, Figure 2(a) illustrates a piece of the HTML page extracted from a specific curriculum vitae. While some unrelated fields are grouped by the same HTML markup, there are freely formatted fields that need additional parsing (e.g., the researchers' address). This fact, along with the existence of optional fields, represents a major challenge for correctly obtaining the information blocks.

For addressing the optional fields, we have crafted a number of regular expressions for discovering and grouping related information into blocks. The regular expressions are dynamically generated and processed by a finite state machine. Also, the expressions are based on the tag names within the HTML file. Most of these blocks roughly correspond to sections of a curriculum vitae, such as affiliation, education, publications, professional history, etc. For handling freely formated fields, we have adopted some heuristics to extract the largest possible amount of information from them. Extracted blocks are then stored in the *CiênciaBrasil* Repository (a relational database) for further processing (Arrow 4 in Figure 1). From the 4,676 Lattes curricula vitae processed in this first step, it extracted 338,968 information blocks with publication data.

In the second step, it gets information blocks (Arrow 5 in Figure 1) and extracts detailed data (Arrow 6 in Figure 1) in order to build the research social network. We could build such a network based on advisor-advisee relationship [Wang et.al 2010], affiliation [Beauchesne 2011], coauthorship [Lopes et al. 2010] and so on. Given the wide range of areas covered by the INCTs, we decided that coauthorship would be the best relationship to build the network (since some of the researchers did their PhDs in international institutions and have changed affiliations over time). Therefore, *CiênciaBrasil* network is based on coauthorship relationships derived from publication metadata extracted from the corresponding information blocks.

As an example, consider the information block shown in Figure 2(b) which contains a publication citation previously extracted from an actual Lattes curriculum vitae. Figure 2(c) shows the result of the second step of the extraction process. This step employs ONDUX [Cortez et al. 2010], an on-demand unsupervised method for text extraction that relies on a knowledge base to perform the extraction process. We chose to use ONDUX because it achieves good extraction results without any user intervention and easily adapts to scenarios in which the input text does not have any standardization.

```
<img src="images/curriculo/ico_doi.gif"></a>Martins, Waister ;
<a href="http://lattes.cnpq.br/3457219624656691" target="blank"
class="coautor">Gon\c{c}alves, Marcos</a> ; Laender, Alberto H. F. ;
<a href="http://lattes.cnpq.br/3527197809276361" target="blank"
class="coautor">Ziviani, Nivio</a> . Assessing the quality of scientific
conferences based on bibliographic citations. Scientometrics <sup>
<img id="01389130_1" src="/buscatextual/images/curriculo/jcr.gif"
class="tipAjax" tooltip="&lt;span&gt;SCIENTOMETRICS (PRINT)
(ISSN 0138-9130) &lt;BR/&gt;Fator de Impacto (JCR 2009): 2.167
&lt;/span&gt;" style=""></sup>, v. 83, p. 133-155, 2010.
```

(a)

Martins, Waister ; Gonçalves, Marcos; Laender, Alberto H. F. ; Ziviani, Nivio . Assessing the quality of scientific conferences based on bibliographic citations. Scientometrics , v. 83, p. 133-155, 2010.

(b)

Title: Assessing the quality of scientific conferences based on bibliographic citations
Authors: Martins, Waister ; Gonçalves, Marcos ; Laender, Alberto H. F. ; Ziviani, Nivio
Venue: Scientometrics
Year: 2010

(c)

Figure 2. Publication metadata extraction

3.3. Data Deduplication

As *CiênciaBrasil* processes data coming from individual curricula vitae (Arrow 7 in Figure 1), each coauthored publication is likely to be listed in more than one Lattes curricula vitae. However, not all of these occurrences are exact duplicates. For instance, author names may appear in different order or differently abbreviated. Moreover, in some cases, even the publication title might be differently written: besides having typos, it is very common to change the paper title between acceptance and publication, so one author may add the original title while another uses the new one. Not considering such duplicates may interfere with derived statistics and further analysis carried out with the stored data.

Hence, it is necessary to carry out a *deduplication* process over all publication entries. We adopt an unsupervised heuristics-based method [Borges et al. 2011] (developed in the context of InWeb). Such a method discards duplicates in which authors' initials and year diverge, thus avoiding unnecessary comparisons. Furthermore, to perform efficiently, the data deduplication process is parallelized using a map-reduce framework[10], which allows new machines to be easily added to improve time performance.

Notice that *nearly-duplicated* publications represent a great challenge for deduplication. In most cases, they consist of publications that share most of the authors, have similar titles and were published in or close to the same year. For example, the article "Infectious bronchitis virus isolate *IBV/Brasil/200/1981* nucleocapsid protein (N) gene" and "Infectious bronchitis virus isolate *IBV/Brasil/PM1/1987* nucleocapsid protein (N) gene" have the same list of coauthors and were published in the same year. That is a simple example that characterizes nearly-duplicated publications as an open problem and on-going work in our research group.

[10]Hadoop Pig: http://pig.apache.org

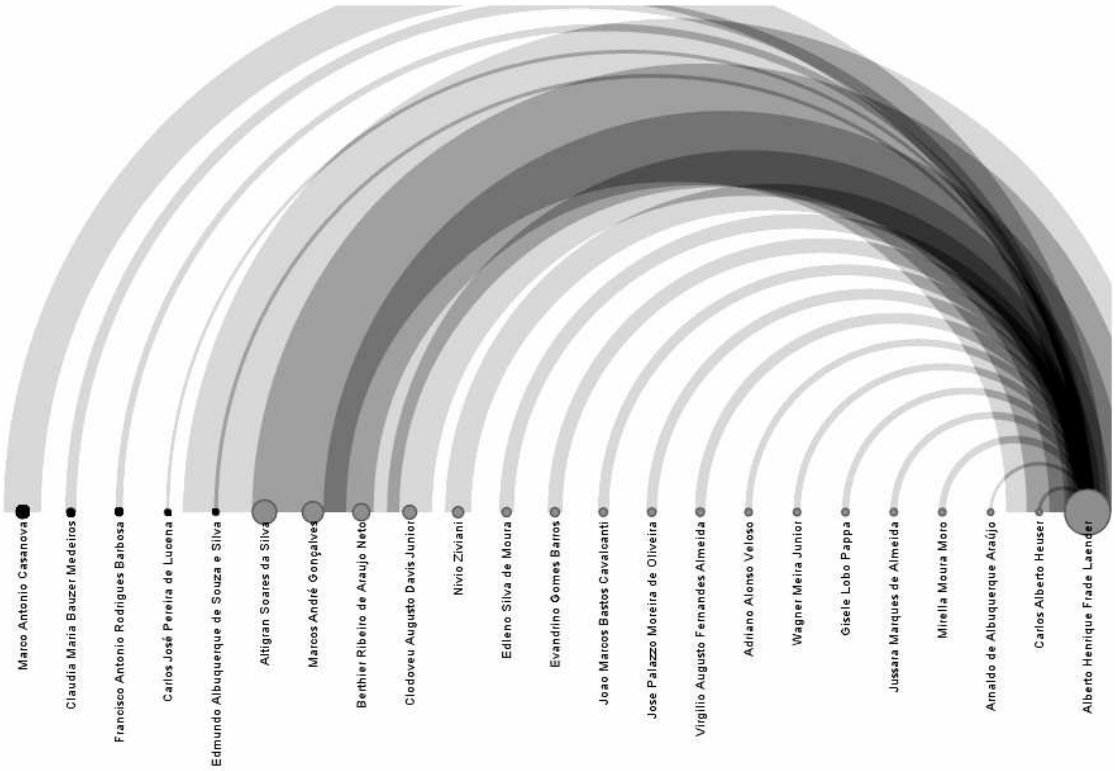


Figure 3. Author graph example

For assessing the effectiveness of the deduplication process, we have analyzed three random subsets of data with 1,507 publications, in which replicas were manually identified. This process has achieved 92% of precision (few false positives have been found) and 100% of recall (all replicas have been found). Therefore, the method from [Borges et al. 2011] does indeed offers an effective solution for handling duplicates on *CiênciaBrasil.* The data deduplication process was able to identify 288,867 unique citations of publications authored by INCT researchers. These unique citations were then stored back into the *CiênciaBrasil* Repository.

3.4. Network Materialization

After the deduplication process, the collected data is ready for materializing the network. This is done by deriving coauthorship relationships based on the existence of common publications in the researchers' curriculum vitae (Arrows 9 and 10 in Figure 1). An effective way to visualize such a network is through graphs. In particular, two different graphs are built: one on authors and one on collaborations within an INCT. Author graphs are built with the perspective of an *individual* author, as illustrated in Figure 3. This way, each researcher that has collaborated with such an author is represented by a colored circle and positioned in a horizontal line (darker circles represent coautho from a different INCT). An arc represents a collaboration between different authors and its thickness represents the strength of this collaboration (number of publications with both of them as coauthors). It is very important to notice that, for building this graph, disambiguation is not a problem since each connection in the graph relies on the existence of a same publication on the curricula vitae of the two related researchers.

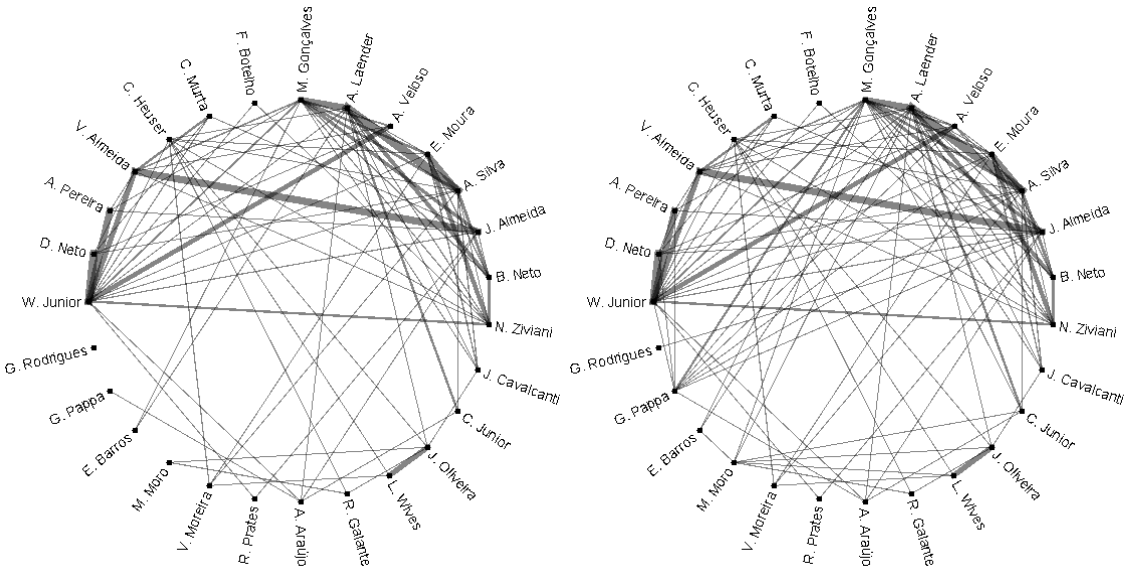


Figure 4. INCT collaboration graph evolution from 2008 (left) to 2010 (right)

The INCT graph shows the collaboration between researchers from the same INCT, as illustrated in Figure 4. Similar to the author graph, one arc represents the collaboration between two authors and its thickness the strength of this collaboration. In this kind of graph, the researchers are presented as points arranged in a wheel format. This network visualization provides an interesting view of coauthorship evolution through time as we can see in Figure 4 by comparing the density of the graphs from 2008 and 2010. On the website, those two graphs are shown as a single one in which blue edges indicate intensified collaborations and yellow edges new collaborations in the period. We are currently working on presenting yearly graphs on the website. Also, the wheel graph style is the same used in an InWeb related work [Lopes et al. 2010].

4. *CiênciaBrasil* Features

The CiênciaBrasil front-end was developed as a web application and written in Django, a Python framework for agile web development[11]. The basic list of features includes the initial webpage, browsing and visualization functions for exploring researchers and INCTs, and research keyword search box. Specifically, the *initial webpage* has explanatory texts about the project's goal, members, context and so on, as illustrated in Figure 5. This page also presents the date when data was last collected as well as the size of the current repository (i.e., number of INCTs, researchers and publications covered).

The user can search for a researcher's name and view details of her curriculum vitae as long as summarized data about her academic production. For example, Figure 6 illustrates the types of publication that a given researcher has in two granularities: by year and total. Specifically, the top part has a line chart for the annual production separated by type (i.e., event papers, book chapters, journal articles, magazine articles and others). The bottom part has a pie chart for the whole production. The researcher page has also a tag cloud based on the titles of her production (note that the clouds are further explained

[11]Django http://www.djangoproject.com

Figure 5. *CiênciaBrasil* main interface

in the next paragraphs). It is very important to notice that the charts are all interactive, i.e. the user can emphasize different parts of the graphs by clicking on them. All charts were created using Highcharts[12], a charting library written in pure JavaScript that allows to add interactive charts in a web application.

The user can also search for an INCT's name and view its detailed information, which includes general information (i.e., INCT coordinator and associated institutions), some statistics and tag clouds. The statistics include a graphic on the percentage of publication types per year. For each year, it sums up the publications of all researchers (within the INCT) grouped by category (or type). The *y* axis shows the percentage and the *x* axis the year. The goal of such statistics is to synthesize the amount and type of the researchers' publications, which makes easier to analyze and evaluate their production.

With such visualizations, it is also possible to compare not only the INCT's productions against themselves but also against areas. For example, Figure 7 shows the accumulated publications grouped by types for two different INCTs: one in software engineering and another in social studies. Notice that with such visualizations, we

[12]Highcharts http://www.highcharts.com/

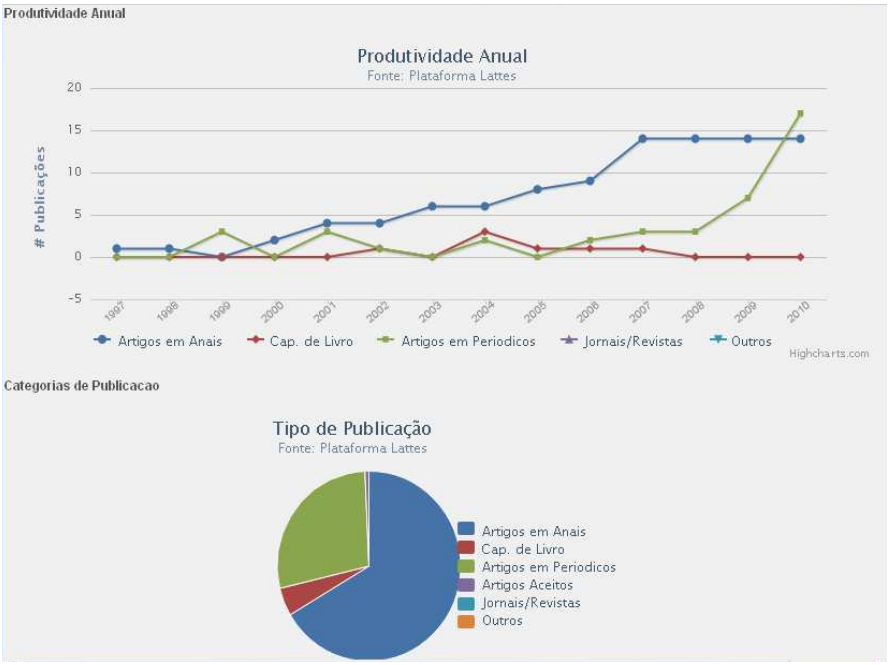


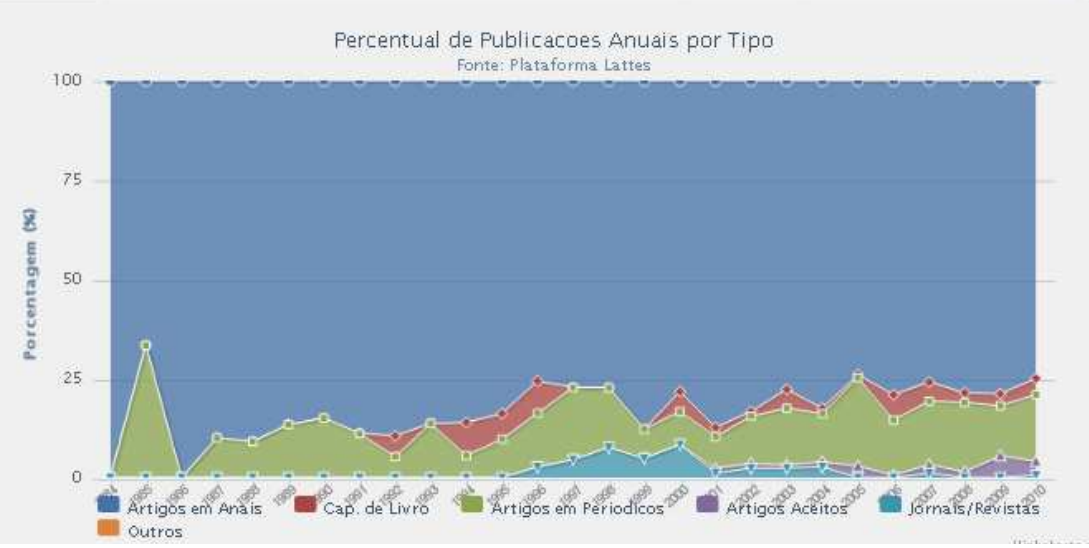
Figure 6. *CiênciaBrasil* detailed statistics on a researcher

can (clearly) see that most of the software engineering publications are on conferences, while the social ones are on journals and book chapters. This once more corroborates the findings from [Laender et.al 2008, Menezes et al. 2009], which shows (using other techniques) that papers on conferences and similar events dominates Computer Science publications.

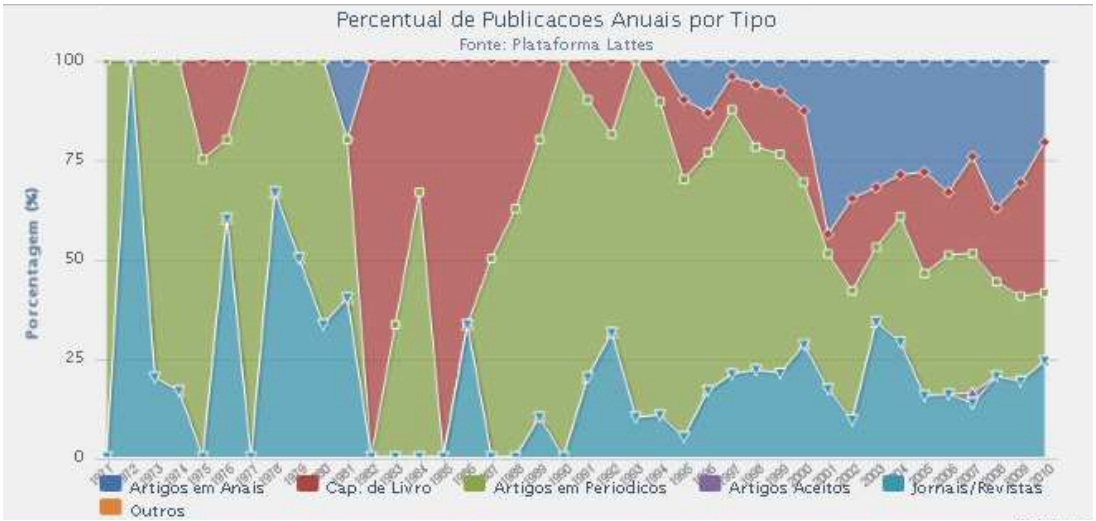
A different perspective about the publications of INCTs and researchers is provided by tag clouds. A tag cloud depicts important terms that characterize a document (often a web page). In general, a tag is usually a single word and its importance to a document defines the size or the color in which it appears in the cloud. However, in scientific documents, it is very common to have expressions besides single words. For example, in a document, you may have the single words “software” and “engineering”, as well as the expression “software engineering”, which has a totally different meaning from its single word components. Needless to say, such expressions are not found only in Computer Science, but in practically all fields. Take for example “homo sapiens”, “sweat equity”, “CT scanning”, “habeas corpus”, “relativity theory” and many others.

In our case, we extract tags from titles of scientific production by applying a number of well known feature-selection heuristics. Such heuristics are based not only on the frequency of an expression but also on it representativeness for a particular researcher or INCT [Koutrika et al. 2009]. In order to allow identifying multi-word expressions we also extract noun phrases and n-grams.

For example, Figure 8 illustrates the tag cloud for the titles of publications authored by members of the INCT on Integrated Studies of the Amazon Biodiversity. The cloud shows words for cities and states covered by the Amazon rainforest (e.g., Amazonas, Brazilian Amazon, estado do Amazonas, Rio Branco, Roraima), species, order, genus, class and family of fauna and flora that are studied over there (e.g., anura,



(a) INCT for Software Engineering



(b) INCT for Studies about the USA

Figure 7. *CiênciaBrasil* statistics on the publications of two different INCTs

apidae, arachnida, bactris gasipaes, Solanum sessiliflorum Dunal) as well as other specific expressions such as “floresta de terra firme” e “abelhas com ferrão”.

Finally, a general keyword search box is available at the top of every page. Its result is a list of both INCTs and researchers whose names approximately match the given keyword. This feature is powered by Xapian[13], an open source search engine library written in C++ with binding for many programming languages. In order to integrate Xapian with *CiênciaBrasil*, we used a modular search extension for Django called Haystack[14], which also was used for indexing. Such an index is necessary in order to speed searches up. Specifically, all INCTs and researchers are described by text documents containing their names, and those documents are used as the target data for

[13]Xapian http://xapian.org
[14]Haystack Search http://haystacksearch.org

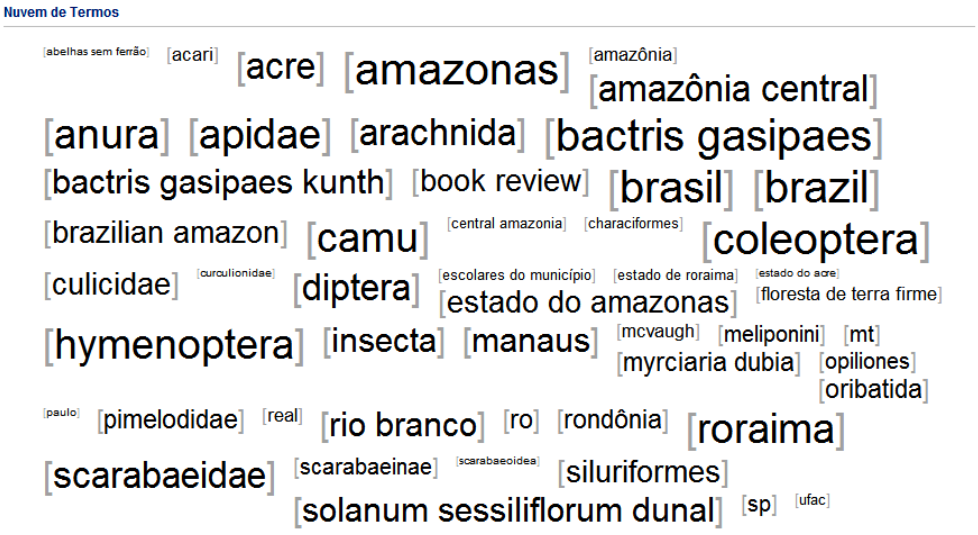


Figure 8. Tag cloud for an INCT on Integrated Studies of the Amazon Biodiversity

indexing. As the index is created just once, we need to perform a re-indexing every time we load new data.

5. Concluding Remarks

Summary. This paper presented an overview of *CiênciaBrasil*, a Web portal for viewing and analyzing a research social network formed by researchers within the Brazilian INCT program. *CiênciaBrasil* has been built from the individual perspective of each INCT researcher by using the Lattes platform as data source. Although the Lattes platform provides a unified and very clean data source for building such a research social network, a number of challenging problems still must be faced. Some of those problems and their solutions are summarized in Table 1.

Future Work. Although *CiênciaBrasil* has been developed as a test platform for experimenting with some of the methods and tools we have developed for web crawling, data extraction and data deduplication within INWeb, we expect that it might also be used as a platform for analyzing and evaluating the results and evolution of the INCT program. To provide more flexibility for such an important task, a major challenge remains: to keep *CiênciaBrasil* up-to-date with Lattes since it is a very dynamic data source. Every day, hundreds of researchers update or create their curricula vitae on Lattes. At the moment, updating the *CiênciaBrasil* Repository requires redoing all the steps described in Figure 1, which is very time consuming. Thus, a more efficient strategy must be devised to maintain *CiênciaBrasil* synchronized as much as possible with the Lattes content.

There are many different ways to improve, explore and visualize the information given by *CiênciaBrasil*. For improving it, we could also consider information about the publications that is available on other publication-oriented websites, such as Google Scholar[15] and CiteSeer[16]. For example, we could extend the publication analysis to

[15]Google Scholar http://scholar.google.com/
[16]Citeseer http://citeseer.ist.psu.edu/

Table 1. Summary of challenges and their solutions

Challenge	Solution
Gather reliable data from individuals for building the research network	Lattes platform was employed as data source
Collecting the specific Lattes curricula (from the set of researchers within INCTs)	Manual input of curriculum vitae's URL followed by a crawling process
Extracting data from specific Lattes fields (since there are no clear delimiters among them)	Regular expressions and ONDUX [Cortez et al. 2010], an on-demand unsupervised method for text extraction
Deduplication of citation records extracted from the individual curricula vitae	An unsupervised heuristics-based method proposed in [Borges et al. 2011]
Searching for groups and individuals on the portal	Xapian, an open source search engine library, and Haystack, a modular search extension for Django
Providing ways of the user to interact with the information	Highcharts, a charting library for interactive charts in a web application
Visualizing the research social network	Building researchers-oriented graphs based on the researchers coauthorships
Visualizing aggregated information on individual and group production	Grouping data for statistics on interactive charts
Building and visualizing tag clouds for word expressions	We extract tags from titles by applying well known feature-selection heuristics.

consider the number of citations of each paper. In the exploration front, we are currently working on adding new social network metrics for providing different analysis. For example, the work on [Lopes et al. 2010] (also developed within InWeb) defines new metrics that could bring some light on different facets of coauthorship relationships. In the visualization front, we are currently working on expanding the views one can get from the network. For example, we plan to add both temporal and geographical dimensions to the author and INCT graphs.

Acknowledgements

This work is partially supported by INWeb (MCT/CNPq grant 57.3871/2008-6) and by the authors' individual grants and scholarships from CNPq, CAPES and FAPEMIG.

References

Beauchesne, O. H. (2011). Map of scientific collaboration between researchers. http://olihb.com/2011/01/23/map-of-scientific-collaboration-between-researchers.

Borges, E. N., Carvalho, M. G., Galante, R., Gonçalves, M. A., and Laender, A. H. F. (2011). An Unsupervised Heuristic-based Approach for Bibliographic Metadata Deduplication. *Information Processing & Management*, Accepted for publication.

Carvalho, M. G., Laender, A. H. F., Gonçalves, M. A., and da Silva, A. S. (2008). Replica identification using genetic programming. In *Procs. of SAC - ACM Symposium on Applied Computing*, pages 1801–1806, Fortaleza, Brazil.

Cortez, E., da Silva, A. S., Gonçalves, M. A., and de Moura, E. S. (2010). ONDUX: on-demand unsupervised learning for information extraction. In *Procs. of SIGMOD Conference*, pages 807–818, Indianapolis, USA.

Geer, D. (2008). Reducing the Storage Burden via Data Deduplication. *Computer*, 41(12):15 –17.

Heydon, A. and Najork, M. (1999). Mercator: A Scalable, Extensible Web Crawler. *World Wide Web*, 2(4):219–229.

Koutrika, G., Zadeh, Z. M., and Garcia-Molina, H. (2009). Data clouds: summarizing keyword search results over structured data. In *Procs. of EDBT - Intl. Conf. on Extending Database Technology*, pages 391–402, Saint-Petersburg, Russia.

Laender et.al, A. H. F. (2008). Assessing the research and education quality of the top Brazilian Computer Science graduate programs. *SIGCSE Bulletin*, 40(2):135–145.

Lane, J. (2010). Let's make science metrics more scientific. *Nature*, 464(7288):488–489.

Lopes, G. R., Moro, M. M., Wives, L. K., and de Oliveira, J. P. M. (2010). Cooperative Authorship Social Network. In *Procs. of AMW - Alberto Mendelzon Workshop on Foundations of Databases*, Buenos Aires, Argentina.

Menezes, G. V., Ziviani, N., Laender, A. H. F., and Almeida, V. A. F. (2009). A Geographical Analysis of Knowledge Production in Computer Science. In *Procs. of WWW - International World Wide Web Conference*, pages 1041–1050, Madrid, Spain.

Nie, Z., Zhang, Y., Wen, J.-R., and Ma, W.-Y. (2005). Object-level ranking: bringing order to Web objects. In *Procs. of WWW - International World Wide Web Conference*, pages 567–574, Chiba, Japan.

Sarawagi, S. and Bhamidipaty, A. (2002). Interactive deduplication using active learning. In *Procs. of KDD - ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pages 269–278, Edmonton, Canada.

Tang et.al, J. (2008). ArnetMiner: Extraction and Mining of Academic Social Networks. In *Procs. of KDD - ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pages 990–998, Las Vegas, USA.

Wang et.al, C. (2010). Mining advisor-advisee relationships from research publication networks. In *Procs. of KDD - ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pages 203–212, Washington, DC.