LattesMiner: a Multilingual DSL for Information Extraction from Lattes Platform

Alexandre D. Alves, Horacio H. Yanasse
National Institute for Space Research (INPE)
alexandre.alves@inpe.br, horacio@lac.inpe.br

Nei Y. Soma
Aeronautic Institute of Technology (ITA)
soma@ita.br

Abstract

The Lattes CV system, a curricular information system maintained by CNPq, is the core of the Lattes Platform. This system is undoubtedly the major source of information on Brazilian researchers. This paper describes "LattesMiner", a multilingual domain-specific language for automatic information extraction from Lattes curricula. It is composed by a set of classes written in Java that allows developers to implement their own applications with a high-level abstraction and expression power. LattesMiner can extract data belonging to the Lattes Platform from any individual researcher or group of researchers by its name or given (ID) number. The data extracted can be analyzed and used, for instance, to identify academic social networks, regional competences, profile of groups in different areas of research etc. We illustrate its use with a case study.

Categories and Subject Descriptors D.3.3 [*Programming Languages*]: Language Constructs and Features

General Terms Domain-Specific Language, Lattes Platform

Keywords Domain-Specific Language, Information Extraction, Academic Social Network

1. Introduction

Lattes Platform (LP) is an information system implanted by CNPq (National Council for Scientific and Technological Development) to manage information on science, technology and innovation related to researchers and institutions in Brazil [6]. This platform is undoubtedly the major source of information available on Brazilian researchers, acknowledged in a recent article published in Nature [13]. The article cites the LP as an example of high-quality database.

The LP is maintained by the Brazilian Government and it includes information systems, databases and Web portals. The Lattes CV system, a curricular information system, is the main component of the platform. Currently, the Lattes CV system stores around 2.000.000 curricula of researchers, lecturers, students and professionals from diverse areas of knowledge with actuation in science, technology and innovation.

The Lattes curriculum (Lattes CV) is a document created by the CNPq with the objective of standardizing and centralizing academic, professional and personal information of the Brazilian scientific community. By using the Lattes CV system it is possible to consult these information at any time via Web. The data of each individual curriculum are filled by the professional him/herself and they have been used by agencies in the country to evaluate researchers, projects, graduate programs etc. Hence, the data are continuously updated by the researchers. Furthermore, the scientific community itself monitors the quality and correctness of the information displayed in the system, since the resource allocation is based upon the comparison of the curriculum of the professionals. Therefore, this system has a very high quality information extraction potential.

In the last years, many works were developed using data extracted from LP of researchers of different areas of knowledge. Some of these works analyzed the profile of the Productivity Research Scholarship fellows in areas such as Public Health [4][22], Dentistry [23][5], Medicine [16][14][19] and Chemistry [21]. Further information were also considered, such as gender and region of the researchers [3] or statistical correlation between the productivity of researchers and his/her proficiency in written English [26]. Master dissertations [7], Doctoral thesis [17] and many other works analyze data extracted from LP in their development. A common problem presented in these works is that the curricula and the information extracted had to be obtained manually.

This paper describes "LattesMiner", an internal multilingual DSL (Domain-Specific Language) for automatic information extraction from Lattes curricula. Observe that, despite being public and accessible via Web[1], the access to

SPLASH'11 Workshops, October 23–24, 2011, Portland, Oregon, USA.
Copyright © 2011 ACM 978-1-4503-1183-0/11/10. . . $10.00

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

the curricula in the LP system is restricted. To perform a search for each registered curriculum an alpha-numeric code (CAPTCHA) is required to avoid automatic searches by scripts.

LattesMiner can extract data belonging to the LP from any individual researcher or group of researchers (up to an entire set of them) by its name or given (ID) number. LattesMiner is composed by a set of classes written in Java that allows developers to implement their own applications with a high-level abstraction and expression's power. The extracted data can be analyzed and used, for instance, to identify academic social networks, regional competences, profile of groups in different areas of research and many other features of interest. Currently, LattesMiner is available in Portuguese and English, and it can be easily extended to other languages.

2. Related Work

From the review of the literature we became aware of two tools that allow the extraction of information from Lattes curricula: Lattes Extrator and scriptLattes.

Lattes Extrator was developed by CNPq itself and it is one of the tools that compose the LP. It is accessible via Web[2] with restricted access. Currently, only licensed institutions can extract data directly from Lattes curricula database of CNPq limited to the data of researchers, lecturers, students and collaborators of their own institutes. The information extracted are available in XML files format, defined by the LMPL (Markup Language of Lattes Platform) community and, the institutions can develop routines to import data to their bases. The extractions are made in batches and they can be configured according to the interest and the permissions of each user.

scriptLattes is a script currently developed in Python for extraction and compilation of bibliographical production, students supervised, participation in examination boards, judging committees, and events, collaboration graphs and research map of a group of researchers on the LP [15]. To run the script it is necessary to create an input file in text format containing the identification number and the name of the researchers, among other optional information. The identification number assigned by CNPq contains 16 digits and it is used as an ID for each Lattes curriculum. The construction of the input file can be very laborious in the case of group of researchers, since each researcher's name must be searched first in the LP to obtain its (ID) number. The tool is restricted to the Linux operating system, therefore, to use it in other operating system recompilation and reconfiguration are required. When the pages are generated in HTML/JSP, the user needs a Web server installed and properly configured to execute dynamic pages in Java. scriptLattes generates reports and charts as HTML pages. Also, the use of the data in other applications is more complex.

Therefore, the creation of alternative more friendly methods for extracting data seems to be of interest. To the best of our knowledge, there is no programming library or DSL to extract data from the Lattes curricula. There are others domains where DSLs have been applied sucessfully [11][12] and they served as the basis for the development of LattesMiner.

3. LattesMiner DSL

LattesMiner is part of a larger project called "Unified System of Curricula and Programs: Identification of Academic Networks - SUCUPIRA", financed by CAPES (Coordination for the Improvement of Higher Education Personnel). The SUCUPIRA project aims to be an automated computational public domain tool to assist users in obtaining performance indicators for lectures, researchers and graduate programs.

LattesMiner is an internal multilingual DSL for automatic information extraction and identification of academic social networks from LP. It is composed by a set of classes written in Java that allows developers to implement their own applications with a high-level abstraction and expression power. LattesMiner allows to extract data belonging to the LP from any individual researcher or group of researchers (up to an entire set of them) by its name or given (ID) number. The extracted data can be analyzed and used, for instance, to identify academic social networks, regional competences, profile of groups in different areas etc.

In the design of LattesMiner DSL the first goal was to define the terms of the problem domain [25]. It is worth mentioning that the Lattes CV is available in Portuguese and in English. Also, the Lattes CV is already being used in other countries of different languages. This was taken into account and LattesMiner was designed to be a multilingual DSL. Currently, LattesMiner can be used in Portuguese and English.

LattesMiner consists of six main components: Data Discovery, Data Acquisition, Data Extraction, Data Structure, Data Visualization and Data Analysis. The output of a component is used as input of another component. An overview of the architecture of library components is illustrated in Figure 1.

The first component, Data Discovery, is used to find the (ID) number of the researchers. Each Lattes CV has an associated URL that allows direct access to it. Usually, only the name of the researcher is available and with the Lattes CV system one cannot perform this search sequentially for any quantity of names since there is a limitation of access. The URL is composed by numbers with 16 digits[3]. An alternative form to access a Lattes CV is using the code of the researcher. LattesMiner DSL allows access to a Lattes CV using any one of the forms and without access restriction. The result of the Data Discovery can be used as input for the Data Acquisition component, that is responsible for down-

[2] http://lattesextrator.cnpq.br/lattesextrator/

[3] http://lattes.cnpq.br/3107924103069456

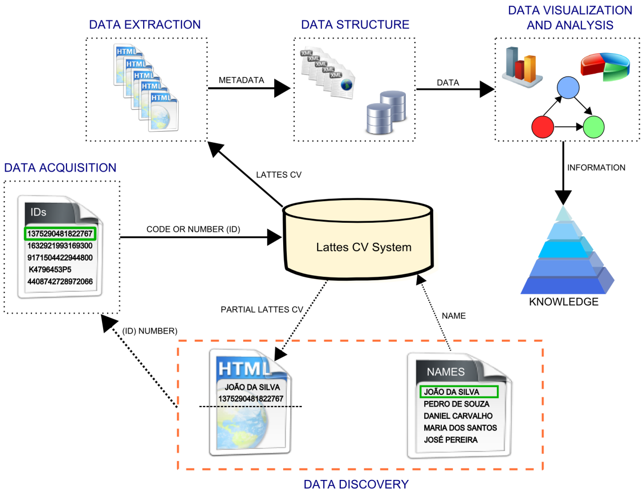


Figure 1: Component architecture of LattesMiner

loading the Lattes curricula of the researchers from Lattes CV system on the Web.

Data Extraction is the main component of LattesMiner DSL. It is responsible for extracting data from the HTML files. Currently, the data that are extracted are shown in Table 1. The technique of infomation extraction based on regular expressions was used. The reason for using regular expressions is because when the Lattes CV is downloaded it is not tag balanced; therefore, it is not possible to use a HTML parser. Also, it was observed that fragments of the Lattes CV in the HTML files have a repetition structure [18][24].

The extracted data can be stored in XML files or in any database using the Data Structure component. In the case of the database, anyone can be used, since LattesMiner DSL has a text file of properties that allows the alteration of such configuration, at any time.

The Data Visualization component is responsible for the identification and visualization of the academic social networks. These networks are identified by checking the relationships between researchers. The Data Analysis component is responsible for the analysis of the data extracted and also for the analysis of the relationships identified. These two last components are under development.

4. Case Study

LattesMiner is composed by a set of classes written in Java and its main class provides the majority of the DSL functionalities. Figure 2 shows a simple UML class diagram describing part of LattesMiner DSL. The `LattesMiner` class is composed by instances of classes `Biodata` and `Board`, in addition to many others not presented here. The class `Biodata`, for example, contains the profile data of the researcher and its corresponding class in the Portuguese language is the class `Perfil`, that is associated to `Biodata` class. The class `BiodataIE` is responsible for extracting data of the researcher and the class `BiodataDao` is resposible for the persistence of such data.

Table 1: Data extracted by the LattesMiner DSL

Biodata	(ID) number, code, name, gender, CNPq grantee of research productivity scholarship level (if applicable), photo, last update date of the curriculum, information of the CV Lattes HTML (date, time and size in KB)
Professional Address	institution, city, state, country, zip code, homepage
Formal Education/ Degree	level, advisor, institution, title, starting and conclusion years, grantee, course, information of the institution (concept, code, acronym and country)
Academic Advisory	level, student, title, institution, course, year of conclusion, type of orientation (advisor or co advisor)
Participation in Examination Boards	type, student's name, title, institution, course, year
Articles in Scientific Journals	article title, authors, journal title, DOI, pages, year, volume, series, number, ISSN, one of the most relevant or not
Complete works published in proceedings of conferences	article title, authors, title event, pages, year, volume, DOI, one of the most relevant or not
Areas of Expertise	major area, area, subarea, specialty
Languages	comprehend, speak, read, write
Bibliographic Citation	all forms of citations informed
Contacts	all (ID) number of researchers cited in Lattes CV

LattesMiner is an internal DSL [8] and was created through a fluent interface [9], that provides a compact and yet easy-to-read representation of the domain problem. Fluent interfaces are implemented using the method chaining. Any method in the chain can be called at any time and any number of times [20]. In addition to the method chaining technique, LattesMiner DSL makes use of static factory methods and imports creating a compact, yet readable DSL. All this can be observed in the illustrative examples presented next.

For the following examples researchers of the Computer Science area with CNPq Research Productivity Scholarship (PQ) were considered. The researchers with PQ are classi-

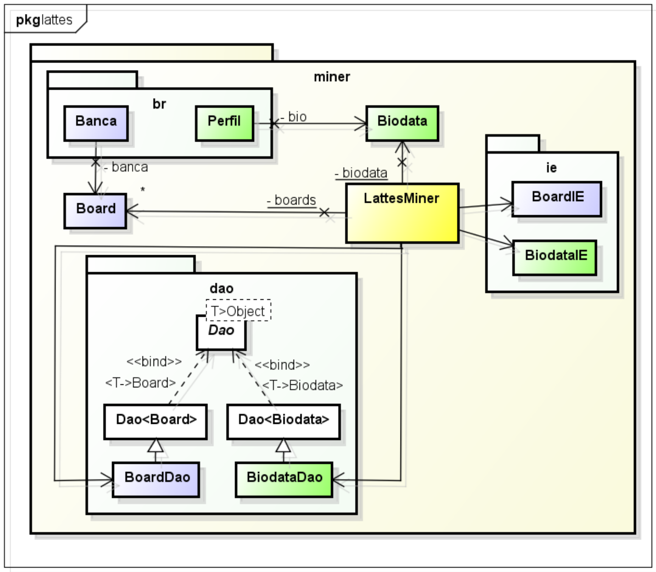


Figure 2: Partial UML class diagram of LattesMiner DSL

fied into six levels: PQ-2F, PQ-2, PQ-1D, PQ-1C, PQ-1B and PQ-1A. The process for choosing if a researcher will receive a scholarship and the level that she(he) is classified is based on scientific and technical merit of both her(his) project and her(his) academic-technical career, and the judgment is by peer review. A list containing all the names of the researchers by area is available in CNPq's site[4]. However, their corresponding (ID) numbers are not provided at this location and it is necessary further processing to discovery them.

The list of names contained in CNPQs page was obtained in August 7, 2011. Only those listed with an indication of being in "Em folha de pagamento" were considered fellows with scholarships. Others, for example, with grants suspended, were not considered. The total number of fellows with scholarships was 376 and the great majority is in category 2 (67.02%), as show in Table 2.

Table 2: Distribution of the CNPq PQ Scholarship in Computer Science by category

Category	n	%
1A	14	3.72
1B	15	3.99
1C	32	8.51
1D	59	15.69
2	252	67.02
2F	4	1.07
Total	376	100.00

The first step is to obtain the names of these researchers and to store them in a text file. In this case the file "names.txt" was created, containing each name in a separate line. The next step is to find the (ID) number of the researchers. The **Listing 1** is the Java application code used to discover the (ID) number of the researchers.

Listing 1

```
import java.util.*;
import lattes.util.Util;
import static lattes.miner.LattesMiner.*;

public class Listing1 {

 public static void main(String[] args) {
  List<String> list = new ArrayList<String>();

   for (String name : Util.getList("names.txt"))
    list.add(search(name));

  Util.setList(list, "ids.txt");
 }
}
```

The `search()` method performs a search by the name of the researcher in the Lattes CV system. If it is found, it returns the (ID) number of the researcher. Otherwise, it returns the name of the researcher. In cases where more than one curriculum with the same name is found, all numbers concatenated and separated by commas are returned. So, it is possible to verify in the file generated if a problem occurred. The result is stored in a text file named "ids.txt". The code previously given corresponds to the Data Discovery component. To the Computer Science area a list containing all the names was found and 13 researchers with homonyms were identified. For example, the researcher "Carlos Eduardo Pereira" has other 15 homonyms registered in LP.

The **Listing 2** shows the code fragment used to download the Lattes curricula of the researchers. It corresponds to the Data Acquisition component. The generated list of (ID) numbers is read and the Lattes CV is downloaded. The Lattes curricula are stored as HTML files and the filename is saved together with the (ID) number of the researcher. The `dir()` method defines the directory where the files are stored. If the directory does not exist, it is automatically created.

Listing 2

```
dir("cvs");
for (String id : Util.getList("ids.txt"))
 download(id).save();
```

After executing these steps it is possible to extract data from the Lattes curricula, as shows in **Listing 3**. Again the generated list of (ID) numbers is read and each HTML file is loaded as a string using the `load()` method. Only part of the data obtained by the suggested DSL was illustrated here due to space limitations and, the code fragment is part of the Data Extraction component. In this illustration, the profile data of the researchers are extracted, together with his

[4] http://www.cnpq.br/

publications in journals and the data of his/her professional address. The method `publications()` can extract publications in proceedings (to do this just use the `CONFERENCE` constant). It is also possible to extract all publications, by using the method without any argument.

Listing 3

```
props("mysql");
for (String id : Util.getList("ids.txt")) {
 load(id).biodata().publications(JOURNAL);
 address().save();
}
```

The `save()` method stores all the data extracted in the database defined in a file of properties (for example, mysql.properties, that can be defined using method `props()`), independently of the order in which the extraction methods are called. Another possibility is to store the data in a XML file. In this case, the method `xml()` is used and each Lattes CV is stored with the corresponding (ID) number of the researcher. These methods are part of the Data Structure component.

The **Listing 4** shows a code fragment to illustrate how the LattesMiner DSL is used to extract information in Lattes CV in different languages, in this case, Portuguese and English; in the first part, in Portuguese, how to get the name of all the students that the researchers examined in examination boards in 2010, and in second part, in English, how to get the name of all the students that the researchers examined in 2010, but limited to doctoral examination boards.

Listing 4

```
for (String id : Util.getList("ids.txt")) {

 // Portuguese
 for (Banca b : carregar(id).bancas().getBancas()) {
  if (b.ano() == 2010)
   System.out.println(b.aluno());
 }

 // English
 for (Board b : load(id).boards().getBoards()) {
  if (b.type() == 'D' && b.year() == 2010)
   System.out.println(b.student());
 }

}
```

The main advantage of LattesMiner DSL in being multilingual is the flexibility offered to the user. Although the conceptual redundancy should be avoided [10], in this case it was necessary because the Lattes CV can be made available both in Portuguese and in English. On the other hand, had the LattesMiner DSL been available only in one language, another guideline “Adopt existing notations domain experts use”, also cited in [10], is not being considered.

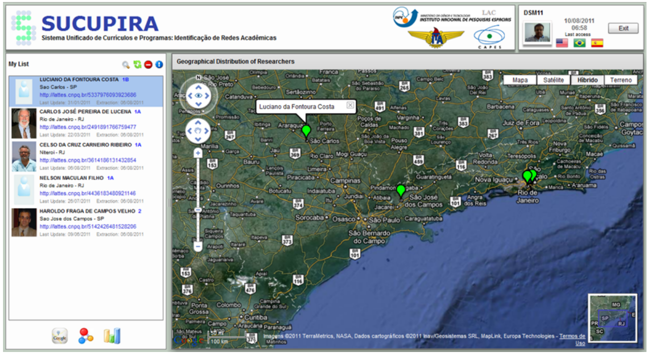
5. Results

In this section, results of a simulated illustrative study are presented. Five researchers from Brazilian Computer Science area (see Table 3) that have published more papers in scientific journals (just the quantity, without any consideration of their quality) were picked. These data were obtained from the database generated by **Listing 3**, using a simple SQL command.

Table 3: Five researchers that have published more in scientific journals

Name	Institution	Level	Total
Luciano da Fontoura Costa	USP	1B	176
Carlos José Pereira de Lucena	PUC-Rio	1A	118
Celso da Cruz Carneiro Ribeiro	UFF	1A	107
Nelson Maculan Filho	UFRJ	1A	107
Haroldo Fraga de Campos Velho	INPE	2	97

Using the LattesMiner DSL, the SUCUPIRA system [1] was developed by the authors of this article. The SUCUPIRA is a system for identification and visualization of academic social networks. Figure 3 shows an initial page of the SUCUPIRA system, emphasizing the geographical distribution of these five researchers. It is possible to visualize in the map where these researchers are working, based on the professional address indicated in the curriculum of each researcher. It can be observed that all the five researchers are from the southeast region, concentrating in Sao Paulo and Rio de Janeiro states.


SUCUPIRA
Sistema Unificado de Currículos e Programas: Identificação de Redes Acadêmicas
My List
Geographical Distribution of Researchers
Mapa
Satélite
Híbrido
Terreno
Luciano da Fontoura Costa
Exit

LUCIANO DA FONTOURA COSTA 1B
Sao Carlos - SP
http://lattes.cnpq.br/5337976093923686
Last Update: 31/01/2011 Extraction: 06/08/2011

CARLOS JOSÉ PEREIRA DE LUCENA 1A
Rio de Janeiro - RJ
http://lattes.cnpq.br/2491891766759477
Last Update: 22/03/2011 Extraction: 06/08/2011

CELSO DA CRUZ CARNEIRO RIBEIRO 1A
Niteroi - RJ
http://lattes.cnpq.br/3614186131432854
Last Update: 06/08/2011 Extraction: 06/08/2011

NELSON MACULAN FILHO 1A
Rio de Janeiro - RJ
http://lattes.cnpq.br/4436183480921146
Last Update: 26/07/2011 Extraction: 06/08/2011

HAROLDO FRAGA DE CAMPOS VELHO 2
Sao Jose dos Campos - SP
http://lattes.cnpq.br/5142426481528206
Last Update: 09/06/2011 Extraction: 06/08/2011

Figure 3: Initial page of the SUCUPIRA system

In Figure 4 the graph of contacts of the five researchers is presented. This graph is defined by the researchers contacts (links to other Lattes CV) contained in their Lattes CV. Every contact contains the (ID) number of the researcher, identifying the relationships between them. Thus, the graph depicts an academic social network of the five researchers. In this network the nodes are presented with a label with the name of the researcher and the colors of the edges represent the number of relationships among researchers, where intensity of the color reflects the number of relationships. The vertices are colored according to the classification level of the scholarship: the color blue indicates level 1A, the color light green indicates level 1B, the color yellow indicates level 1C, the color orange indicates level 1D and the color red indicates level 2. The black color is used to represent the researchers that do not belong to the Computer Science group that is being analyzed.

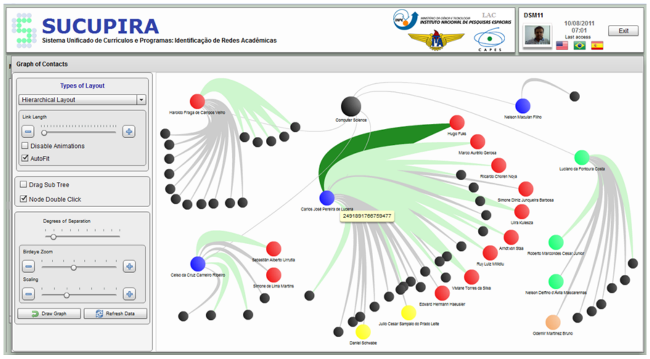


Figure 4: Graph of contacts of the five researchers that have published more in scientific journals

In this academic social network it is possible to visualize the relationships between the five researchers with a degree of separation equal to 2. In this network, it is clear that two of the researchers have no contact with any of the others 375 researchers in the Computer Science area. On the other hand, the researcher “Carlos José Pereira de Lucena” has 11 contacts, being 9 of them of the category 2 and 2 of the category 1C. The “main” relationship of this researcher is with the researcher “Hugo Fuks”, which is highlighted by the green edge.

This was just an illustration of many other possibilities of knowledge discovery that may be carried out using the LattesMiner DSL.

6. Conclusions and Future Work

Currently, the Lattes CV are available in HTML format. This imposes a further effort to information extraction. LattesMiner DSL however does not depend on the data format of the Lattes CV because it allows users to program their own applications with a high-level abstraction. If the data format is eventually modified, the DSL interface remains the same. An advantage of LattesMiner DSL compared to Lattes Extrator and scriptLattes is that it searches by the name of the researcher while Lattes Extrator and scriptLattes only allow the searches by the (ID) number of the researcher. In addition, LattesMiner is multilingual; it can be used with different languages. Another advantage of LattesMiner is that the data extracted are stored in a structured format (XML or database), allowing these data to be easily used by other applications.

LattesMiner DSL is already being successfully used to develop the SUCUPIRA system [1] and it has already been used to analyze the profile of the Productivity Research Scholarship Fellows in the areas of Production Engineering and Transportation of CNPq [2], in less than one hour. A beta version of LattesMiner will be available soon for testing and it will be free to users and developers. The use of the language is very simple, just the library “LattesMiner.jar” has to be imported and the library to the database (*e.g.* “mysql-connector-java-5.1.8-bin.jar”) if the user wish to store the data in a database.

The future step that is already being implemented in the LattesMiner DSL is a statistical analysis of the data.

Acknowledgments

The authors acknowledge the financial support of CAPES and CNPq.

References

[1] A. D. Alves, H. H. Yanasse, and N. Y. Soma. 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)*, pages 371–376, Chaves, Portugal, 06 2011.

[2] A. D. Alves, H. H. Yanasse, and N. Y. Soma. Perfil dos bolsistas pq das áreas de engenharia de produção e de transportes do cnpq: enfoque na subárea de pesquisa operacional. In *XLIII Simpósio Brasileiro de Pesquisa Operacional*, Ubatuba, SP, 08 2011.

[3] D. Arruda, F. Bezerra, V. Neris, P. Rocha De Toro, and J. Wainera. Brazilian computer science research: Gender and regional distributions. *Scientometrics*, 79:651–665, 2009. ISSN 0138–9130. URL `http://dx.doi.org/10.1007/s11192-007-1944-0`.

[4] R. B. Barata and M. Goldbaum. Perfil dos pesquisadores com bolsa de produtividade em pesquisa do cnpq da Área de saúde coletiva. *Cadernos de Saúde Pública*, 19:1863–1876, 12 2003. ISSN 0102–311X.

[5] R. A. Cavalcante, D. R. Barbosa, P. R. F. Bonan, M. B. de Oliveira Pires, and H. Martelli-Júnior. Perfil dos pesquisadores da Área de odontologia no conselho nacional de desenvolvimento científico e tecnológico (cnpq). *Revista Brasileira de Epidemiologia*, 11:106–113, 03 2008. ISSN 1415–790X.

[6] CNPq. Plataforma lattes. `http://lattes.cnpq.br/`, 2011.

[7] F. de Simone Cividanes. Collectlattes : Sistema para extração do conhecimento sobre a plataforma lattes. Dissertação (mestrado em engenharia eletrônica e computação), Instituto Tecnológico de Aeronáutica (ITA), São José dos Campos, 2010.

[8] M. Fowler. A pedagogical framework for domain-specific languages. *IEEE Software*, 26(4):13–14, 2009. ISSN 0740–7459. doi: http://doi.ieeecomputersociety.org/10.1109/MS.2009.85.

[9] M. Fowler. *Domain-Specific Languages*. Addison-Wesley Professional, 2010.

[10] G. Karsai, H. Krahn, C. Pinkernell, B. Rumpe, M. Schindler, and S. Vőlkel. Design guidelines for domain specific lan-

guages. In *The 9th OOPSLA Workshop on Domain-Specific Modeling*, Orlando, USA, 10 2009.

[11] A. A. Kejriwal and M. Bedekar. Mobidsl - a domain specific langauge for mobile web applications: developing applications for mobile platform without web programming. In *The 9th OOPSLA Workshop on Domain-Specific Modeling*, Orlando, USA, 10 2009.

[12] S. H. Khandkar and F. Maurer. A domain specific language to define gestures for multi-touch applications. In *10th SPLASH Workshop on Domain-Specific Modeling (DSM'10)*, Reno/Tahoe, USA, 10 2010.

[13] J. Lane. Let's make science metrics more scientific. *Nature*, 464(7288):488–489, 03 2010. ISSN 1476–4687. URL `http://dx.doi.org/10.1038/464488a`.

[14] H. Martelli-Júnior, D. R. B. Martelli, I. G. Quirino, M. C. L. A. Oliveira, L. S. Lima, and E. A. de Oliveira. Pesquisadores do cnpq na Área de medicina: comparação das áreas de atuação. *Revista da Associação Médica Brasileira*, 56:478–483, 2010. ISSN 0104-4230.

[15] J. P. Mena-Chalco and R. M. C. Junior. scriptlattes: an open-source knowledge extraction system from the lattes platform. *Journal of the Brazilian Computer Society*, 15(4):31–39, 2009. ISSN 0104–6500.

[16] P. H. C. Mendes, D. R. B. Martelli, W. P. de Souza, S. Q. Filho, and H. Martelli-Júnior. Perfil dos pesquisadores bolsistas de produtividade científica em medicina no cnpq, brasil. *Revista Brasileira de Educação Médica*, 34:535–541, 12 2010. ISSN 0100–5502.

[17] M. L. Moreira. Formação de competências em ciência e tecnologia espaciais: Uma análise da trajetória da pós-graduação no instituto nacional de pesquisas espaciais. Tese (doutorado em política científica e tecnológica), Universidade Estadual de Campinas (Unicamp), Campinas, 2009.

[18] T. Nanno, S. Saito, and M. Okumura. Structuring web pages based on repetition of elements. In *Second International Workshop on Web Document Analysis*, Japão, 2003.

[19] E. A. Oliveira, R. Pecoits-Filho, I. G. Quirino, M. C. Oliveira, D. R. Martelli, L. S. Lima, and H. Martelli-Júnior. Perfil e produção científica dos pesquisadores do cnpq nas Áreas de nefrologia e urologia. *Jornal Brasileiro de Nefrologia*, 33:31–37, 03 2011. ISSN 0101-2800.

[20] A. Ruiz and J. Bay. An approach to internal domain-specific languages in java. `http://www.infoq.com/articles/internal-dsls-java`, 2008.

[21] N. C. F. Santos, L. F. de Oliveira Cândido, and C. L. Kuppens. Produtividade em pesquisa do cnpq: análise do perfil dos pesquisadores da química. *Química Nova*, 33:489–495, 2010. ISSN 0100-4042.

[22] S. M. C. Santos, L. S. Lima, D. R. B. Martelli, and H. Martelli-Júnior. Perfil dos pesquisadores da saúde coletiva no conselho nacional de desenvolvimento científico e tecnológico. *Physis: Revista de Saúde Coletiva*, 19:761–775, 2009. ISSN 0103–7331.

[23] A. C. Scarpelli, F. Sardenberg, D. Goursand, S. M. Paiva, and I. A. Pordeus. Academic trajectories of dental researchers receiving cnpq's productivity grants. *Brazilian Dental Journal*, 19:252–256, 2008. ISSN 0103–6440.

[24] S. Vadrevu, F. Gelgi, and H. Davulcu. Information extraction from web pages using presentation regularities and domain knowledge. *World Wide Web*, 10(2):157–179, 06 2007. ISSN 1386-145X. doi: http://dx.doi.org/10.1007/s11280-007-0021-1.

[25] A. van Deursen, K. Paul, and V. Joost. Domain-specific languages: an annotated bibliography. *ACM SIGPLAN Notices*, 35(6):26–36, 2000. ISSN 0362-1340. doi: http://doi.acm.org/10.1145/352029.352035.

[26] S. Vasconcelos, M. Sorenson, and J. Leta. A new input indicator for the assessment of science & technology research? *Scientometrics*, 80:217–230, 2009. ISSN 0138–9130. URL `http://dx.doi.org/10.1007/s11192-008-2082-z`.