DEPARTAMENTO DE ENGENHARIA ELÉTRICA

Internal Research Reports

ISSN

Number 34 | December 2013

# Getting to Know the Contents of Bibliographic References of ETDs

Ana Maria Beltran Pavani

PUC RIO

**Internal Research Reports**

Number 34 | December 2013

# Getting to Know the Contents of Bibliographic References of ETDs

Ana Maria Beltran Pavani

**CREDITS**

**Publisher:**
**MAXWELL / LAMBDA-DEE**
**Sistema Maxwell / Laboratório de Automação de Museus, Bibliotecas Digitais e Arquivos**
**http://www.maxwell.vrac.puc-rio.br/**

**Organizers:**
Alexandre Street de Aguiar
Delberis Araújo Lima

**Cover:**
Ana Cristina Costa Ribeiro

# GETTING TO KNOW THE CONTENTS OF BIBLIOGRAPHIC REFERENCES OF ETDs

***Ana M B Pavani***
apavani@lambda.ele.puc-rio.br
Pontifícia Universidade Católica do Rio de Janeiro
Rio de Janeiro, Brazil

**Abstract**

ETDs contain precious information – state-of-the-art bibliographic reviews. Graduate students starting their research can browse the references and learn about important works in their areas of interest. At the same time, references can be viewed from a different stand point – as holders of information on:

- Relations there exist among ETDs and other works of their authors, their advisors and other faculty;
- Types of works and languages they are written in;
- Different profiles of types of works and languages among the disciplines;
- Evolution of types and languages as time goes by, in general and for different disciplines.

This addresses the first part of a project in progress focusing the contents of references in ETDs. The project has the objectives of: (1) Identifying ETDs that are references in ETDs and trying to establish a measure analogous to the impact factor; (2) Mapping endogeny in graduate programs; (3) Analyzing the evolution of reference types from traditional (paper) to online works.

In order to fulfill the objectives, a set of tools was developed. They are devoted to ETDs and not to theses & dissertations in general because they are applied to digital files; the reason is operational.

**Keywords:** ETD, bibliographic references; impact factor; bibliographies

## 1. INTRODUCTION

Garfield (2005) states that he introduced the idea of an impact factor in an article he published in 1955 (Garfield 1955). Later, in the 1960s the journal impact factor was created to evaluate, compare and select journals.

In the five decades from this first article, the impact factor has been used and also, according to the creator (Garfield 1955), has "become so controversial. Like nuclear energy, the impact factor is a mixed blessing." At the same time, it is an index that is used to evaluate journal and, as a consequence, authors and even institutions.

In the case of theses & dissertations, in Brazil, the "quality" of the journal in which the resulting papers are published is important not only for the work, but for the graduate program. As a matter of fact, CAPES – Coordenação de Aperfeiçoamento de Pessoal do Ensino Superior (http://www.capes.gov.br/), the federal agency in charge of the assessment of all graduate programs in the country, has a list called Qualis with international and national journals and the grading points they assign to graduate program evaluation. The Qualis list will be addressed later because it was used in this part of the project.

A lot of energy has been devoted to discuss the strengths and weaknesses of the impact factor and how it is to be viewed for different types of works in journals. An example is the work of Amim & Mabe (2000), where differences in behavior of the index are presented when different areas of knowledge, types and sizes of journals are present; the time frame of the measurement is also shown to be a cause of varying results. Some disciplines do not have journals in the extent of others, like biomedical areas and science & technology do. So, alternative impact factors can be searched in other types of works. Another example of discussion of the journal impact factor is presented by Dong, Loh & Mondry (2005) who also analyzed the

impact factor and how the way calculations are performed may bias results, and presented questions the impact factor can not answer and alternative measures to evaluate journal impact.

Different types of articles and journals show different behaviors related to impact factors (Amim & Mable 2000). An interesting result is presented by Garfield (1994) concerning review articles that are more cited because they contain extensive bibliographies.

This characteristic is common to ETDs which contain state of the art bibliographic reviews. During many training sessions, the participants – librarians, graduate program deans and graduate students – described theses & dissertations as a source of bibliographies.

But theses & dissertations are neither articles nor journals. The current concept of impact factor can not directly be applied to them. Theses & dissertations are presented in all areas in which graduate programs are available regardless the culture of journal publishing.

This work addresses the preliminary results of a project that is in progress.

## 2. CONTEXT OF THE PROJECT

PUC-Rio – Pontifícia Universidade Católica do Rio de Janeiro has had a digital library since 1995, when the Maxwell System (http://www.maxwell.lambda.ele.puc-rio.br/) was created. In 1998, PUC-Rio registered the system in INPI – Instituto Nacional de Propriedade Industrial (http://www.inpi.gov.br/), the Brazilian patent office.

The system started as a digital library of courseware and soon became a general purpose digital library. If a current term is used, it became an institutional repository. This was a consequence of the university using the system to store and deliver other types of documents.

During the second semester of 1999, the ETD module of the system was developed and the first ETD was published in May 2000. The first ETDs came from three graduate programs that made them a requirement – Civil Engineering, Business Administration and Electrical Engineering. In 2001, other graduate programs joined the ETD project with local requirements. In August 2002, ETDs started being required for all graduate programs.

Currently, there are over 2,400 ETDs on the system. The university's production of theses and dissertations is between 500 and 600 per year. Specific situations lead to the current number of ETDs. Civil Engineering had collected digital files of all theses and dissertations since 1997. Electrical Engineering started retrospective digitization of the whole collection. In other graduate programs, some digital files were obtained by person-to-person contact with alumni.

The ETD project has been enhanced by adding new functions that are not related to allowing access to ETDs. They focus on data and information focusing graduate program administrators. Two sets of functions were implemented in the last 3 years – the first set contained statistics on the production of ETDs and the second on accesses to the works in the collection.

Analysis of bibliographic references is the third enhancement to PUC-Rio's ETD project.

## 3. THE PROJECT

ETDs are theses and dissertations in digital formats. Worldwide, most of them are texts; all ETDs on the Maxwell System are texts in pdf format. The fact that ETDs are in digital formats makes them suitable for the analysis of references since ICT – Information and Communication Technology tools can be used.

This project has the objectives of: (1) Identify ETDs that are references in ETDs and try to establish a measure analogous to the impact factor; (2) Map endogeny in graduate programs; (3) Analyze the evolution of reference types from traditional (paper) to online works.

The objectives required that references be extracted from the texts, classified, identified, stored on the database and analyzed. The project, therefore, started with the development of a set of computational tools to perform these operations.

This paper presents the tools and some results from a sample set of ETDs. It is divided in sections to deal with the following topics:

- Extract references from ETDs – references are written in XML with a DTD for each type according to the ABNT – Associação Brasileira de Normas Técnicas (http://www.abnt.org.br/) standard used at PUC-Rio. Each set of references is identified so that the ETD is known;
- Develop a set of tools to analyze quantitatively the references – numbers and averages by type, language, graduate program, year, etc. Currently, there are 9 types of analysis;
- Develop a set of tools to analyze qualitatively the references – relations of ETDs with works of authors, supervisors, other students of the same supervisor, other faculty, etc. Currently, there are 6 types of analysis;
- Choose a set of ETDs to be the sample and apply the tools to analyze them;
- Discuss the preliminary results in terms of reference contents;
- Discuss the problems with references in ETDs on the system;
- Present the next steps to automate extraction, identification, storage and classification of the references;
- Discuss with graduate programs faculty to gather suggestions for further developments and for a possible index (IF for T&D?).

## 4. REFERENCES: FROM ETDs TO THE DATABASE

The first activity was to get references from ETDs. All ETDs have a 'references' section in a pdf file. References should be written according to the Brazilian reference bibliographic standard established by ABNT – Associação Brasileira de Normas Técnicas (http://www.abnt.org.br/). Unfortunately, there are many ETDs that are not compliant to the standard; this will be discussed in a later section.

References are written as XML records with tags that identify the elements. There are different sets of tags according to the type of reference – books, chapters of books, articles in scientific journals, conference proceedings, websites, etc. Elements are dived in two sets:

- DCMES – Dublin Core Metadata Element Set elements – all elements that belong to DCMES (http://www.dublincore.org/) are identified as <dc.name of the element> and </dc.name of the element>;
- All other elements are identified as <maxwell.name of the element> and </maxwell.name of the element>. This set has more elements since there are many types of references and each one requires a special set of attributes for its description, though some elements are common to almost all (title and language are two examples).

All XML records are stored on table fields of the database and are managed by IBM DB2 Text Extender (http://www-306.ibm.com/software/data/db2/extenders/text/). Different types of search are used. Other types of data about ETDs and their references are stored on the database for regular relational database manipulation; the database management system is IBM DB2 (http://www.ibm.com/software/data/db2/udb/).

Records needed to be created to prove concept. The first solution was manual cataloging using a program that generates XML records from a user friendly template. A similar solution will be used when students start using the templates to generate the XML records when they submit their ETDs files.

At the same time, there are more than 2,400 ETDs on the system and their references must be captured as well as all others that are submitted before students are required to generate them.

To solve the problem of current ETDs, a second solution is currently under final test. It is based on tools developed by Mr. Akeo Tanabe, M Sc, a researcher in the Computer Science Department of PUC-Rio. He works with components in computational processes in the domain of natural languages for Brazilian Portuguese. This solution extracts records from reference sections of ETDs and, automatically, classifies and identifies them. Results are to be reviewed and corrected by a human being since the non compliance to the standard yields some errors. Mr. Tanabe is adjusting his solution and the number of errors has decreased in the last tests. The complete workflow of this process must be established given that ETDs grow at a rate of 500 to 600 per year.

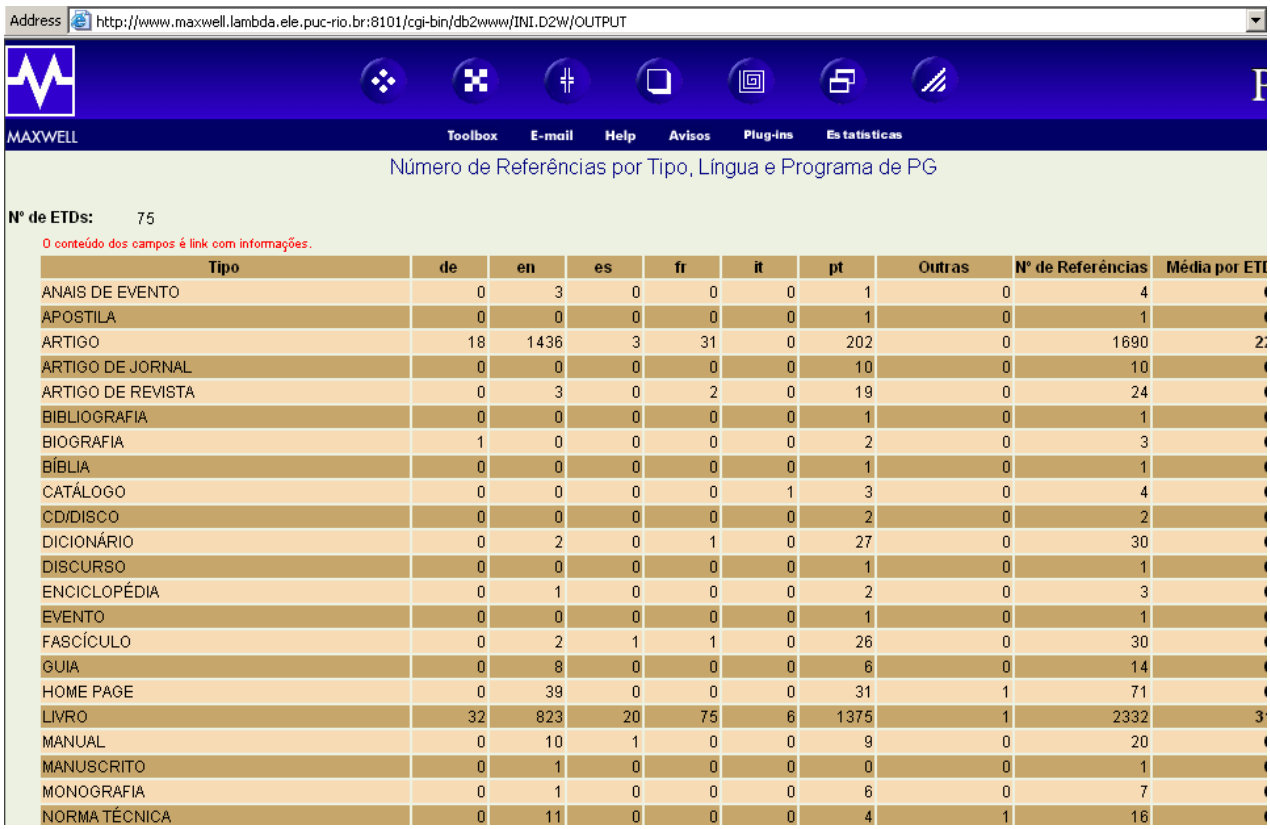
## 5. ANALYSIS OF THE REFERENCES

As mentioned in section 3, two types of analysis are performed with reference records:

- Quantitative analysis

Qualitative analysis has the objective of yielding information on numbers related to references by type, language and year for each ETD and for graduate programs. Currently, there are 9 different types of combinations of these attributes to compute results. Some outputs present time-series of the results while others show total numbers (all years).

Figures 01 and 02 show the outputs for the number of references by type, language and graduate program.


Address http://www.maxwell.lambda.ele.puc-rio.br:8101/cgi-bin/db2www/INI.D2W/OUTPUT

MAXWELL Toolbox E-mail Help Avisos Plug-ins Estatísticas

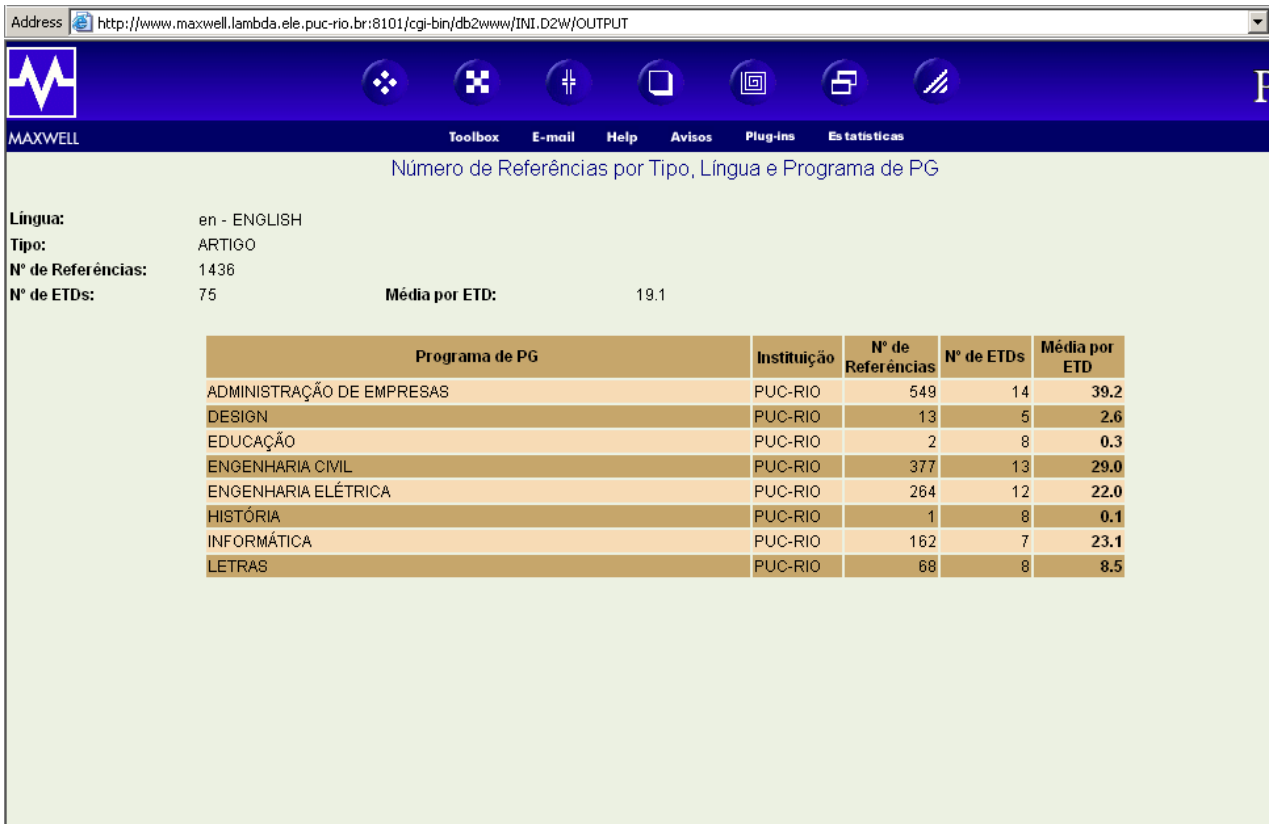
Número de Referências por Tipo, Língua e Programa de PG

**Nº de ETDs:** 75

O conteúdo dos campos é link com informações.

| Tipo | de | en | es | fr | it | pt | Outras | Nº de Referências | Média por ETD |
|---|---|---|---|---|---|---|---|---|---|
| ANAIS DE EVENTO | 0 | 3 | 0 | 0 | 0 | 1 | 0 | 4 | |
| APOSTILA | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 1 | |
| ARTIGO | 18 | 1436 | 3 | 31 | 0 | 202 | 0 | 1690 | 2 |
| ARTIGO DE JORNAL | 0 | 0 | 0 | 0 | 0 | 10 | 0 | 10 | |
| ARTIGO DE REVISTA | 0 | 3 | 0 | 2 | 0 | 19 | 0 | 24 | |
| BIBLIOGRAFIA | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 1 | |
| BIOGRAFIA | 1 | 0 | 0 | 0 | 0 | 2 | 0 | 3 | |
| BÍBLIA | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 1 | |
| CATÁLOGO | 0 | 0 | 0 | 0 | 1 | 3 | 0 | 4 | |
| CD/DISCO | 0 | 0 | 0 | 0 | 0 | 2 | 0 | 2 | |
| DICIONÁRIO | 0 | 2 | 0 | 1 | 0 | 27 | 0 | 30 | |
| DISCURSO | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 1 | |
| ENCICLOPÉDIA | 0 | 1 | 0 | 0 | 0 | 2 | 0 | 3 | |
| EVENTO | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 1 | |
| FASCÍCULO | 0 | 2 | 1 | 1 | 0 | 26 | 0 | 30 | |
| GUIA | 0 | 8 | 0 | 0 | 0 | 6 | 0 | 14 | |
| HOME PAGE | 0 | 39 | 0 | 0 | 0 | 31 | 1 | 71 | |
| LIVRO | 32 | 823 | 20 | 75 | 6 | 1375 | 1 | 2332 | 3 |
| MANUAL | 0 | 10 | 1 | 0 | 0 | 9 | 0 | 20 | |
| MANUSCRITO | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 1 | |
| MONOGRAFIA | 0 | 1 | 0 | 0 | 0 | 6 | 0 | 7 | |
| NORMA TÉCNICA | 0 | 11 | 0 | 0 | 0 | 4 | 1 | 16 | |

Figure 01 – First part of the output for references by type, language and graduate program – type and language are presented and table cells are sensitive




| Programa de PG | Instituição | Nº de Referências | Nº de ETDs | Média por ETD |
|---|---|---|---|---|
| ADMINISTRAÇÃO DE EMPRESAS | PUC-RIO | 549 | 14 | **39.2** |
| DESIGN | PUC-RIO | 13 | 5 | **2.6** |
| EDUCAÇÃO | PUC-RIO | 2 | 8 | **0.3** |
| ENGENHARIA CIVIL | PUC-RIO | 377 | 13 | **29.0** |
| ENGENHARIA ELÉTRICA | PUC-RIO | 264 | 12 | **22.0** |
| HISTÓRIA | PUC-RIO | 1 | 8 | **0.1** |
| INFORMÁTICA | PUC-RIO | 162 | 7 | **23.1** |
| LETRAS | PUC-RIO | 68 | 8 | **8.5** |

Figure 02 – Second part of the output for references by type, language and graduate program – after clicking the cell for articles in English in scientific journals the numbers by graduate program are shown

- Qualitative analysis

Qualitative analysis has the objective of yielding information, not necessarily represented by numbers, that takes into account relations among works, authors, faculty, graduate programs and journals in the Qualis list.

Currently, there are 8 types of searches and matches. One shows the numbers of each type of reference in an ETD, for any ETD, and shows the references when the ETD is public. Dealing with restricted works was a concern because the references, as parts of the work, are not shown when the author does not allow public access to a thesis or dissertation. Others identify references whose authors are related to the ETD: the supervisor; the ETD author; other faculty of the graduate program; other students of the same advisor; other faculty of the university; any author in the authority table. The last type is the one that finds ETDs whose references are articles in a chosen Qualis listed journal.

Figures 03, 04 and 05 show the outputs for references whose author is an advisor, after the advisor was selected. Figure 05 shows the situation of a public ETD, whose references can be displayed. When the ETD is restricted, figure 04 is the last possible information – only the numbers of references per type.

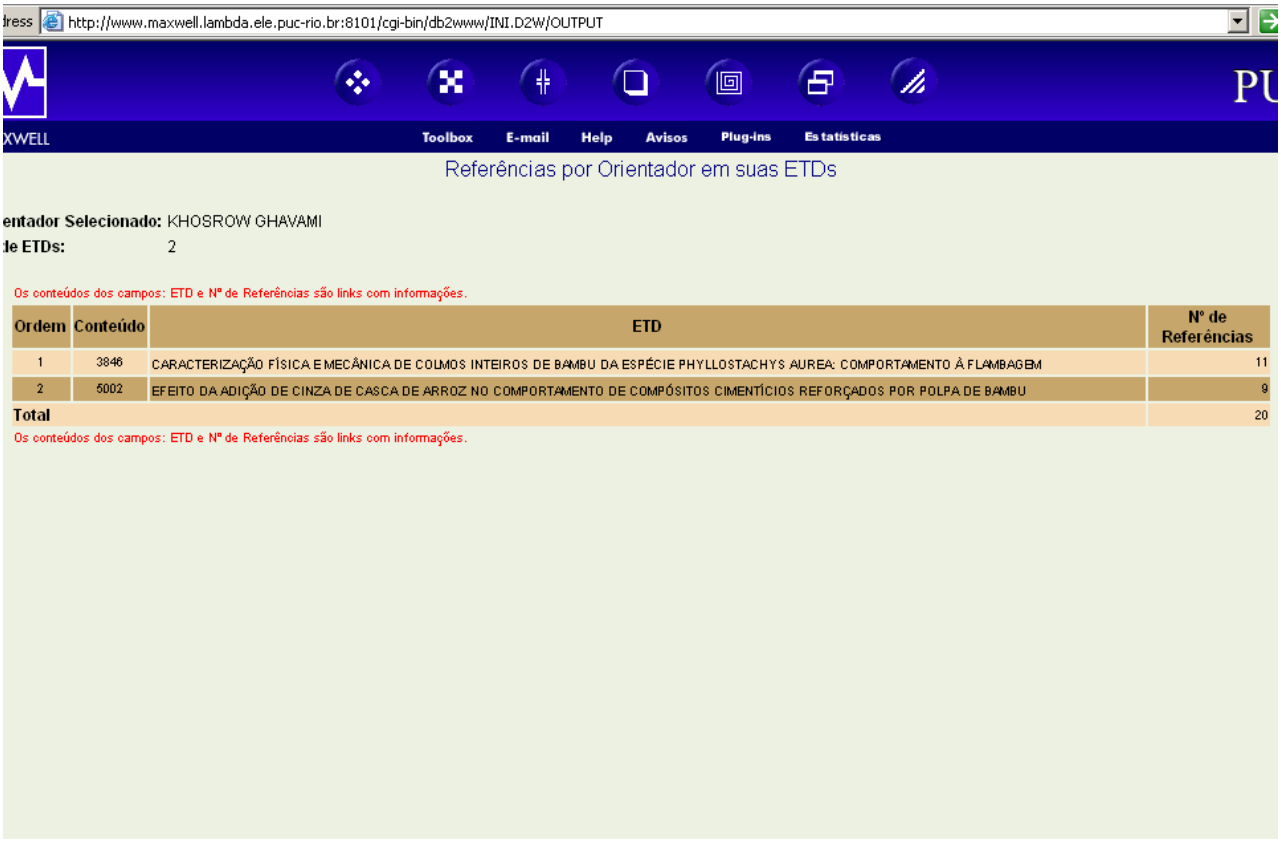



Figure 03 – First part of the output for an advisor in his/her ETDs – cells with numbers of references are sensitive

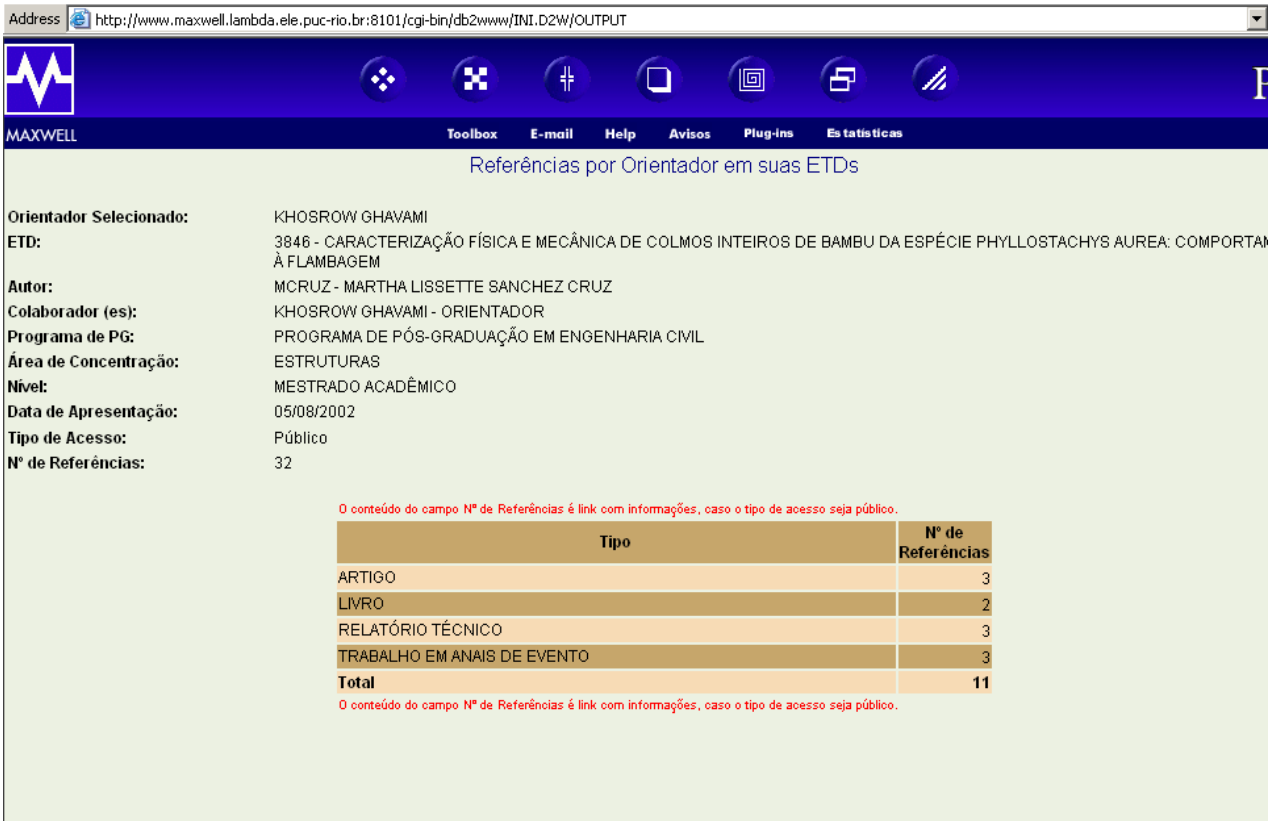



Figure 04 – Second part of the output for an advisor ETDs after choosing the ETD – cells with numbers of references are sensitive if the ETD is public


Figure 05 – Third part of the output for an advisor ETDs after choosing the ETD – the references whose author is the advisor are shown

Altogether, there are 17 options of viewing references. New programs may become necessary as the project continues.

## 6. SAMPLE SET TO TEST THE ANALYSIS OF THE REFERENCES

There are over 2,400 ETDs on the system from all graduate programs of PUC-Rio. Some programs have ETDs that go back further in time than 2002, so their time-series are longer. This may happen because the program started collecting digital files before others (for example Civil Engineering) or because the administrators decided to digitize paper theses and dissertations (for example Electrical Engineering) or because they contacted some alumni to get the digital files (for example Business Administration).

In order to start the project, a sample set was chosen according to the following criteria:

- ETDs should come from the 3 centers of the university;
- Both the master and the doctoral levels should be included;
- Graduate programs with the longest time-series on the system should have the preference;
- Public ETDs should preferably be chosen over restricted ones.

Currently, there are 5,608 references from 75 ETDs coming from both master and doctoral levels from 8 graduate programs. The ETD distribution by center and program is:

- Human Sciences and Theology – Design (5 M), Education (4 M and 4 D) and Languages (4 M and 4 D);
- Social Sciences – Business Administration (9 M and 5 D) and History (4 M and 4 D);
- Science & Technology – Computer Science (3 M and 4 D), Civil Engineering (7 M and 6 D) and Electrical Engineering (6 M and 6 D).

It is planned that all ETDs on the data base be examined and also this process be applied when new ETDs are cataloged and uploaded on the system. The numbers will grow fast since the process of extracting references is under automation.

## 7. RESULTS FROM THE SAMPLE SET

The sample set is small if compared to the collection – approximately 3.1%. But even such small sample yielded interesting results. Some are in terms of languages, others in terms of types.

- Languages

Languages are examined in 3 groups, in spite the existence of the numbers by language. The groups are English (en), Portuguese (pt) and other languages (ot). Other languages in this work are French, German, Italian, Spanish and others that were not accounted individually in the system. Table 01 shows the numbers of references in each group, by graduate program, and the corresponding averages.

There are 3,076 references of all types in English in the sample set. The averages by graduate program are presented in table 01.

| | Ref | ETDs | Ave | Ref (en) | Aver (en) | Ref (pt) | Aver (pt) | Ref (ot) | Aver (ot) |
|---|---|---|---|---|---|---|---|---|---|
| **Business Admin** | 1,642 | 14 | 117.3 | 1,117 | 79.8 | 464 | 33.1 | 64 | 4.6 |
| **Civil Engineering** | 780 | 13 | 60.0 | 649 | 49.9 | 107 | 8.2 | 20 | 1.5 |
| **Computer Science** | 429 | 7 | 61.3 | 387 | 55.3 | 41 | 5.9 | 1 | 0.1 |
| **Design** | 373 | 5 | 74.6 | 68 | 13.6 | 296 | 59.2 | 9 | 1.8 |
| **Education** | 529 | 8 | 66.1 | 25 | 3.1 | 481 | 60.1 | 22 | 2.8 |
| **Electrical Eng** | 624 | 12 | 52.0 | 571 | 47.6 | 52 | 4.3 | 0 | 0.0 |
| **History** | 741 | 8 | 92.6 | 45 | 5.6 | 594 | 74.8 | 105 | 13.1 |
| **Languages** | 490 | 8 | 61.3 | 214 | 26.8 | 265 | 33.1 | 11 | 1.4 |

Table 01 – References in English (en), Portuguese (pt) and other languages (ot) in the sample set of 75 ETDs

Table 02 shows statistical information about the averages per language of the sample – mean and median (Barros 2001). The averages per language were chosen because the numbers of ETDs per graduate program were different, ranging from 5 to 14.

| | Mean | Median |
|---|---|---|
| **English** | 35.2 | 37.4 |
| **Portuguese** | 23.2 | 33.1 |
| **Other languages** | 3.2 | 15.50 |

Table 02 – Statistical information about the sample – mean and median for English, Portuguese and other languages in the sample set of 75 ETDs

When English is considered, graduate programs in Business Administration and Science & Technology have averages above the mean and the median. All other graduate programs are below both measures. When Portuguese is considered, graduate programs in Science & technology are below in both measures.

The average number of references per ETD per program is shown in table 01 too. It can easily be seen that Business Administration has the highest average of reference per ETD – it is 27% higher than History the second in the rank.

When the number of ETDs is higher than 75, these trends may vary. Other graduate programs may bring new insights to the language situation, though Portuguese and English are expected to be the most used.

- Types of references

Books, chapters of books, articles in scientific journals, works in conference proceedings and theses & dissertations are the most used references. Table 03 shows the numbers for the sample. The total number of references is 5,608, as mentioned earlier.

| | Number | Percentage |
|---|---|---|
| **Articles in journals** | 1,690 | 30.1 |
| **Books** | 2,332 | 41.6 |
| **Chapters of books** | 564 | 10.1 |
| **Theses & Dissertations** | 221 | 3.9 |
| **Works in Conf Proceedings** | 377 | 6.7 |
| **Other types** | 424 | 7.6 |

Table 03 – Numbers and percentages of references by type in the sample set of 75 ETDs

Table 03 shows that T & D are less than 10% of the references. A curious number related to them is that from the 221, 169 are in Portuguese, 48 in English and the others in other languages.

When books are concerned, Portuguese comes first with 1,375, then English with 823 and other languages account for the remaining number. Articles in journals switch the language pattern; English comes first with 1,436, then Portuguese with 202 and then other languages.

At the moment, automation of the process is expected before additional programs are developed. The results of the analysis with more ETDs can guide the next steps.

## 8. PROBLEMS WITH THE REFERENCES IN ETDs ON THE DATABASE

The choice of starting this part of the project with extraction of references manually proved to be good. It allowed humans to examine the way references are written. The main problems can be grouped in:

- Classification – in many cases the references were so imprecise that it was not possible to assign a type to them – no title of a journal or number of an issue. Google helped to solve many of the problems; other problems were solved by contacting faculty in the graduate program.
- Identification of authors – the authors' names were not written in the ABNT standard, for there were many versions of the same author, specially when the authors had many middle names. For authors who belong to PUC-Rio (students and faculty) an authority table was created; currently it is fed every time an ETD enters the system.
- Time and space – many references did not have years and/or places of publishing. This is an important information when time series are analyzed.
- Identification of the same reference – different students and even the same student, in different parts of the work, identified references in different ways. This was very common with scientific journals that were abbreviated in many different ways. This problem also impacts the matching with journals in the Qualis list.

When the process is automated, and the many tests have proven this, there must be a final review by a human and also there will be items that the system will not be able to identify. The latter is approximately 10% as shown by the initial tests.

## 9. COMMENTS AND NEXT STEPS

Although the results are from a small sample set, some points are interesting. The ones related to language and type are the most remarkable. Others can be mentioned – the way the types are distributed by the graduate programs, the difference between the master and the doctoral levels and the numbers of references whose author is the supervisor when the topic is very specific.

At the moment, the first next step is to automate the process of extraction of references for the current ETDs – besides completing the tests and approving the final version of the extraction software, the workflow must be modelled. A following step is submitting a proposal to the Dean of Graduate Studies to request that students use the template to generate XML records when they deposit their ETDs – there are between 500 and 600 new ETDs per year. The third step is to view the results of the current programs when more ETDs have their references ready to be analyzed. This step will probably show the need of different data manipulation. The fourth step is try to identify an index, analogous to the impact factor, but for T & D. During steps 2, 3 and 4 there will be interviews with advisors and graduate program administrators.

## 10. REFERENCES


**Amim, Mayour & Mabe, Michael**
Impact Factors: Use and Abuse
*Perspectives in Publishing*, No. 1, October 2000
Elsevier
Captured from http://www.elsevier.com/framework_editors/pdfs/Perspectives1.pdf in May 2006

**Barros, Mônica**
*Probabilidade: um Curso Introdutório*
Papel & Virtual
Brasil, 2001

**Dong, Peng; Loh, Marie & Mondry, Adrian**
The "impact factor" revisited
*Biomedical Digital Libraries* 2005, **2**:7 (5 December 2005)
Captured from http://www.bio-diglib.com/articles/browse.asp?volume=2 in May 2006

**Garfield, Eugene**
Citation indexes to science: a new dimension in documentation through association of ideas
*Science* 122(3159): 108 -111, 1955
Captured from http://garfield.library.upenn.edu/essays/v6p468y1983.pdf in May 2006

**Garfield, Eugene**
The ISI impact factor
Originally published in the Current Contents print editions June 20, 1994, when Thomson Scientific was known as The Institute for Scientific Information® (ISI®).
Captured from http://scientific.thomson.com/free/essays/journalcitationreports/impactfactor/ in May 2006

**Garfield, Eugene**
The Agony and the Ecstasy – The History and Meaning of Journal Impact Factor
Talk presented at the *International Congress on Peer Review And Biomedical Publication*
Chicago, USA, September 16, 2005
Captured from http://www.garfield.library.upenn.edu/papers/jifchicago2005.pdf in May 2006


## AKNOWLEDGEMENTS


- This work was partially financed by FAPERJ – Fundação Carlos Chagas Filho de Amparo à Pesquisa do Estado do Rio de Janeiro (http://www.faperj.br/) under grant E-26/170.040/2002.
- The software products used in the project were made available by IBM through the Academic Initiative Program (http://www.ibm.com/university/).
- Mr. Akeo Tanabe (akeo@teccomm.les.inf.puc-rio.br) deserves a special aknowledgment for accepting to test his work with the ETD data..