DEPARTAMENTO DE ENGENHARIA ELÉTRICA

Internal Research Reports

ISSN

Number 26 | October 2012

Comparing Accesses to ETDs and Journals in Education and Languages Available from the Same Repository

Ana Maria Beltran Pavani

PUC RIO

Internal Research Reports

Number 26 | October 2012

Comparing Accesses to ETDs and Journals in Education and Languages Available from the Same Repository

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

This article was originally published in the Proceedings of the 15th International Symposium on Electronic Theses and Dissertations, Peru, September 2012. It is also available from http://www.etd2012.edu.pe/pdf/presentation/APavaniETD2012FT.pdf

Comparing Accesses to ETDs and Journals in Education and Languages Available from the Same Repository

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

Abstract

This work addresses a comparison between accesses to two sets of resources that are made available from the Maxwell System of PUC-Rio. The first set is made of ETDs in the areas of Education and Languages. The second contains six journals – three in Education and three in Languages. The time frame of the comparison is 28 months from Mar 2010 to Jun 2012. The analysis is performed considering accesses from Brazil, the United States, Portuguese-speaking countries, Spanish-speaking countries and all the others. The results seem to indicate that there are no signifficant differences between ETDs and journals and between Education and Languages. They also show that accesses from Brazil are higher than 75% of all accesses.

Keywords: accesses; ETD; journal; Education; Languages;

01. Introduction

Pontifícia Universidade Católica do Rio de Janeiro (PUC-Rio) is a private university in Rio de Janeiro, Brazil. Though it is a new university when the international the scenario is considered, it has a solid set of graduate programs ranked among the best in the country. The graduate programs are in three areas:

- Humanities & Theology
- Science & Technology
- Social Sciences

PUC-Rio has had an ETD program since 2000 and in August 2002 ETDs became mandatory. Currently, the number of ETDs is over 6,200 from 29 graduate programs in the three areas. The Graduate Program in Electrical Engineering has digitized the complete collection of printed T&Ds (theses and dissertations) and is responsible for almost 18% of all ETDs. Two other graduate programs are in the process of digitization – Civil and Mechanical Engineerings. These actions are important beucase they allow the ETD collection to grow at a faster rate than that of defended T&Ds.

ETDs are made available from the Maxwell System (http://www.maxwell.lambda.ele.puc-rio.br/). The same system hosts many other digital resources – senior projects, courseware, articles, journals and books are some examples. Journals are published on this system and they number 11; the first started in Jun 2003. They are in the areas of Humanities & Theology and Social Sciences.

There is a reasonable set of statistics of both production (publishing) and accesses to contents. Both ETDs and journals have many accesses from different parts of the world. Journals have a common characteristic with some subsets of ETDs – they are in the same areas and editors are faculty in the programs. This is the case of the journals selected for this analysis.

Access data started being gathered on Jun 01, 2004 ; data are related to all contents.

This work is a follow on of previous works by Pavani & Mazzeto (2010) and Pavani (2011) that examined accesses to ETDs and came to the conclusion that language was an important factor influencing them. The objective of this work is to examine if ETDs behave differently from other contents in the same areas of knowledge; only contents available from the Maxwell System are considered so that technical conditions are the same. It analyzes access patterns of ETDs and journals in Education and Languages, both in the Humanities & Theology area. Its tries to identify similarities and differences in acesses from different countries and/or language groups.

Education and Languages were chosen because more than 50% of the journals belong to them, while the remaining five journals are in Design, International Relations, Social Work and Theology.

The time frame of the analysis in this work is from Mar 2010 to Jun 2012. This means that data gathered along 28 months are used. It is quite different from the one used in the previous works and the reason for this choice is the availability of a more significant number of journal issues which occurred in the last 30 months.

Section 2 addresses the reasons for choosing the areas of Education and Languages as well as the time frame; it also presents some numbers related to the collection. Section 3 explains how data on accesses are generated. Section 4 brings some data of the previous works as well as current data – computed for this article. Section 5 comments the results.

02. Numbers and Choices – Profile of the Collection

ETDs started being published in 2000 and have been mandatory for 10 years. For this reason it is a large collection and grows at a steady rate; sometimes it speeds up when digitization projects are under way.

At the same time, journals are optional and depend on research groups that decide to publish them. The number of journals available from the Maxwell System is 11. Four are in Languages and three are in Education; all but one are indexed on DOAJ – Directory of Open Access Journals (http://www.doaj.org/). They are:

- Boletim SOCED (Education)
- Educação On-line (Education)
- Fonogramas[1] (Languages)
- Pesquisas em Discurso Pedagógico (Languages)
- Revista escrita[2] (Languages)
- Sociologia da Educação[3] (Education)
- Tradução em Revista (Languages)

[1] The last journal to be published; it started in May 2012 and it is not considered in this analysis due to insufficient data. It is not indexed on DOAJ.
[2] The first journal to be published; it started in Jun 2003.
[3] The last journal to be published that is considered in this work; it started in Mar 2010.

When journals began being published, ETDs had already been available in the two areas.

a. Some Numbers Related to ETDs

The numbers of ETDs in Mar 2010 and in Jun 2012 are shown in table 1.

	# Mar 2010	% Mar 2010	# Jun 2012	% Jun 2012
Humanities & Theology	1,239	24.52	1,454	23.49
Social Sciences	1,072	21.22	1,465	23.66
Science & Technology	2,741	54.26	3,272	52.85
Total	**5,052**	**100.00**	**6,191**	**100.00**

Table 1 – Numbers and percentages of ETDs on the system Mar 2010 – Jun 2012.

When Education and Languages are considered separately the numbers considerabily change; table 2 shows the the numbers.

	# Mar 2010	# Jun 2012	Average[1]
Education	186	229	208.64
Languages	421	486	466.04

Table 2 – Numbers and average numbers of ETDs in Education and Languages on the system Mar 2010 – Jun 2012.

[1] The average is computed along the 28 months of observation.

Table 2 shows that the numbers are quite different for the two graduate programs.

An important information about ETDs is related to language. Though PUC-Rio has allowed ETDs to be published in languages other than Portuguese, the two sets under consideration have 100% of their items in this language. Table 3 shows the languages of PUC-Rio's ETDs in Jun 2012.

	Education	Languages	Complete Collection
English	0.00	0.00	0.62
Portuguese	100.00	100.00	99.36
Spanish	0.00	0.00	0.02

Table 3 – Percentages of ETDs by language in Jun 2012.

Table 3 shows that the percentage of ETDs in Portuguese is almost 100. For practical purposes, it can be considered a collection of ETDs in Portuguese.

Descriptive data on ETDs that are used to map the profiles of the different graduate programs contain: (1) author; (2) supervisor; (3) language; (4) committee; (5) graduate program; (6) area of research; (7) level; (8) some dates; and (9) awards. Different combinations of these data yield the results referred to as Production Statistics. Not all data items are used in the results presented in this work.

b. Some Numbers Related to Journals

As with ETDs, numbers related to journals are quite different in Education and Languages. Table 4 shows the numbers of issues and items (articles, poems, chronicles, reviews, etc) in the journals in each area.

		# Mar 2010	# Jun 2012	Average[(1)]
Education				
	# Issues	13	21	16.86
	# Items	148	223	181.39
Languages				
	# Issues	19	28	23.79
	# Items	265	387	328.75

Table 4 – Numbers and average numbers of journal issues and articles in Education and Languages on the system Mar 2010 – Jun 2012.

[(1)] The average is computed along the 28 months of observation.

Journals allow articles to be published in languages other than Portuguese. Table 5 shows the languages of the articles in journals in Jun 2012.

	Education	Languages	All Journals
English	2.01	2.34	1.50
Portuguese	97.32	95.55	97.78
Spanish	0.67	2.11	0.72

Table 5 – Percentages of articles by language in Jun 2012

The percentage of articles in Portuguese is a little lower than that of ETDs, but it is still a collection in Portuguese.

Descriptive data on journals and articles that are used to map the profiles of the journals contain: (1) author(s); (2) journal; (3) issue; and (4) date. Different combinations of these data yield the results referred to as Production Statistics. As in the case of ETDs, not all data items are used in the results presented in this work.

c. Decision on the Analysis of Data

Tables 2 and 4 show that the sets of ETDs and journals in Education and Languages have different profiles. In both cases Languages offers more contents than Education. Concerning ETDs, the numbers related to Languages are more than twice as many the corresponding numbers in Education.

This fact lead to the decision not to compare accesses in absolute values but to use percentages. At the same time, tables 3 and 5 show that works in Portuguese account for over 95% in all cases.

Production data are used along with access data to yield qualified Access Statistcs – accesses mapped to the areas of the contents generate percentages.

03. Numbers – Accesses to Collection Items

Section 2 presented numbers related to the profiles of the two types of contents under consideration. This section addresses the numbers of accesses they had in the specified time frame.

Access data are computed according to the following steps:

- The Apache Server log is processed to identiffy rows that indicate accesses to contents; this happens once every hour. Identification means: (1) the content ID on the system; and (2) the country (from an IP x country international table) where accesses came from.
- The result of the processing is included on a table of the database that stores numbers of accesses to each accessed content from each country in a year-month. This means that this table constains a row for each content-country-year-month and the last datum of the row is the corresponding number of accesses.

This procedure yields data that combined with Production Statistics are the source of the results that known as Access Statistics. Access Statistics qualify when and where accesses came from as well as the sets that the accessed items belong to. An example of set is journal and issue, a second example is graduate program and level.

04. Numbers – Results

a. Previous Results

As in the previous works, countries were divided in groups. Two groups are of paramount importance for the analysis – Portuguese and Spanish speaking countries. The reason for such an importance lies on the fact that over 99% of all items on the Maxwell System are in Portuguese and that both languages are quite similar, specially when scholarly communication is under consideration and potential readers are educated persons. Tables 3 and 5 in the previous section showed the percentages of works by language.

The groups of countries are:

- Brazil – home country of the collection
- United States – analyzed separately due to the size, HDI – Human Development Index and number of speakers of both languages in the population
- Pt-speaking countries – Portuguese speaking countries(1)
- Es-speaking countries – Spanish speaking countries(2)
- Other countries

(1) Countries that have Portuguese as one of their official languages: Angola, Cape Verde, East Timor, Equatorial Guinea, Guinea-Bissau, Macau (Special Administrative Region of the People's Republic of China), Mozambique, Portugal and São Tomé and Príncipe.
(2) Countries that have Spanish as one of their official languages: Argentina, Bolivia, Chile, Colombia, Costa Rica, Cuba, Dominican Republic, Ecuador, El Salvador, Guatemala, Honduras, Mexico, Nicaragua, Panama, Paraguay, Peru, Puerto Rico (Commonwealth of Puerto Rico), Spain, Uruguay and Venezuela.

Pt- and Es-speaking countries are quite different. In order to evaluate potencial accesses from these two groups of countries, Pavani & Mazzeto (2010) defined an index – I – that contained information on the sizes of their populations and their standard of living (education and access to the Internet are a part of it).

$$I = (HDI) \times (Population) \quad (1)$$

I is an index that was created to express a "quantity" that takes into consideration both the population and HDI – Human Development Index. Both populations and HDIs were obtained from UNDP Development Report (2010) published by the UNDP – United Nations Development Program.

The United Nations Development Program created the HDI – Human Development Index (UNDP, 1990). It "introduced a new way of measuring human development by combining indicators of the life expectancy, education attainment and income into a composite human development index (HDI)". The numbers used in 2010 were the ones published in 2009. Pavani (2011) updated the information since new numbers came out in 2010.

Table 6 shows total population, average HDI and I for both groups.

	Es-speaking group	Pt-speaking group
Total population	420,281,000	57,858,800
Average HDI	0.707	0.527
Index I	**309,420,871**	**25,114,111**

Table 6 – Index I for the Es- and Pt-speaking Groups.

When ETDs were considered in the 2010 and 2001 works, it was shown that when Brazil and the United States were not included, the remaining (international) group had accesses that were predominantely from Pt- ans Es-speaking countries. When Pt- and Es-speaking countries were taken apart, accesses from Pt-speaking countries largely exceeded the ones from the other group. And when the Pt-speaking group was analyzed by itself, Portugal had the most accesses though its population is small; HDI made the difference.

b. Current Results

Accesses data have been collected for the 28 months from Mar 2010 to Jun 2012 for both ETDs and journals that were specified in section 2. In reality, they were extracted from the complete data that have been stored from Jun 2004 on. As it was mentioned before, the choice of this time frame was a consequence of the number of journal issues that was not significant before Mar 2010 in the areas under consideration.

- Data on Accesses to ETDs

Since Education and Languages were chose for the analysis, access data were extracted for the two graduate programs for the months from Mar 2010 to Jun 2012. They were computed separately.

At the same time, the whole set of ETDs had their results computed in the same ime frame – to allow comparison of the chosen areas with the complete collection. The results for the three sets were computed percentage wise.

Table 7 shows the results.

	Education	Languages	Complete Collection
Brazil	82.01	79.37	78.47
United States	10.71	8.23	11.85
Pt-speaking Group	3.44	5.67	4.78
Es-speaking Group	0.39	0.68	0.73
Others	3.44	6.06	4.18

Table 7 – Percentages of accesses from different groups of countries to ETDs in Education, Languages and the complete PUC-Rio collection between Mar 2010 and Jun 2012.

Some characteristics of the results shown in table 7 can be commented.

- Brazil accounts for approximately 80% of all accesses to ETDs for the three sets. This is not a surprise.
- Accesses from Brazil to the whole collection are lower than both to Education and Languages – this means that there are other areas whose accesses are even lower than 78.47%.
- Considering the complete collection, the percentage of accesses from all countries except Brazil and the United States is 9.68. This result can be compared to the one presented by Pavani (2011) for the months from Jun 2004 to Jun 2011 – it was 8.48%. There is no signficant difference between the two numbers.
- Though numbers in the same rows in the three coluns are not the same, they are compatible and have the same pattern – Brazil has the highest percentage (78.47 – 82.01), followed by the United States (8.23 – 11.85); Pt-speaking countries and Others are almost tied up and the Es-speaking group comes last.
- If accesses from Brazil and from the United States are not considered, the percentage of accesses from Pt- and Es-speaking countries accounts for more that 50% of all accesses (51.17 – 56.86). This number is lower than 69.03 that was found in the previous work.

- Data on Accesses to Journals

The way data were compiled was the same used in the case of ETDs. Table 8 shows the results.

	Education	Languages	Complete Collection
Brazil	86.48	86.90	86.76
United States	2.22	4.19	3.54
Pt-speaking Group	4.59	3.01	3.53
Es-speaking Group	2.96	1.01	1.66
Others	3.75	4.89	4.51

Table 8 – Percentages of accesses from different groups of countries to journals in Education, Languages and the complete PUC-Rio collection between Mar 2010 and Jun 2012.

Chacteristics of the results shown in table 8 can also be commented.

- Accesses from Brazil show the highest percentages for the three sets. This is not a surprise.
- Accesses from Brazil are almost the same for the three sets. In this case, it is reasonable to expect that other areas have similar percentages of accesses from Brazil.
- Considering the complete collection, the percentage of accesses from all countries except Brazil and the United States is 9.7. This is almost the same number as the one obtained for ETDs in the previous section – 9.68.
- If accesses from Brazil and from the United States are not considered, the percentage of accesses from Pt- and Es-speaking countries accounts for more than 65% of all accesses in Education but for only a little over 45% in Languages. This is a significant difference. For the complete collection, the corresponding percentage is 53.51.

- Comparing Access Data – ETDs and Journals

The most visible points to observe are:

- In both cases accesses from Brazil show the highest percentages, but in the case of journals the percentages are higher than those of ETDs.
- Percentages of accesses from the United States are significantly lower in the case of journals than in the case of ETDs. The difference is highest in Education 10.71% x 2.22%.
- Accesses from the Es-speaking group are higher in the case of journals than in the case of ETDs.
- Accesses from 'Others' are not significantly different in the two cases.

It is important to remark that the characteristics observed in Education and Languages are similar to the ones of the whole collection.

There is a point, though, that must be emphasized – table 1 shows that more than 50% of all ETDs are in Science & Technology. Accesses from the United States show their highest percentage for the complete collection. It is worthed examining the characteristics of accesses to ETDs in Science & Technology for the same time frame. There are no journals in Science and Technology, it is not possible to compare to journals.

05. Comments

The first comment is the most obvious – most accesses come from Brazil and in both cases it is a very high percentage. At the same time, ETDs seem to be 'more international'. The complete collection of ETDs is the 'most international' of the six sets that were examined. Is this caused by the ETDs in Science & Technology? This is a topic to study. Would this signficantly change if the percentages of ETDs and articles in English were higher? This is a second topic to study – currently the percentages are very low as shown in tables 3 and 5. Though access numbers are available by ETD and by article, the number of such items in other languages is not significant to allow comparisons.

A second comment derives from metadata availbility for harvesting. The Maxwell System is an OAI-PMH (http://www.openarchives.org) data provider. It has two separate URLs to serve metadata. The first (http://www.maxwell.lambda.ele.puc-rio.br/ibict.php) is dedicated to ETDs and it provides metadata in

two formats – mtd2-br and oai-dc; the first is the Brazilian national format used by institutions that are members of BDTD – Biblioteca Digital de Teses e Dissertações (http://bdtd.ibict.br/), the Brazilian national consortium. The second URL (http://www.maxwell.lambda.ele.puc-rio.br/DC_Todos.php) serves metadata of all items available from the system; the format is oai-dc. So there is no difference between ETDs and journals concerning providing metadata to be harvested.

A comment related to indexing on other systems is important. Both ETDs and journals are indexed by Google. So in this case, there is not difference between ETDs and journals either. Journals are indexed on DOAJ and are linked to from the websites of the departments that publish them. ETDs are made available from many union catalogs that feed on metadata from both BDTD and NDLTD – Networked Digital Library of Theses and Dissertations (http://www.ndltd.org/), the international consortium.

At a first glance it seems reasonable to infere that if international accesses are to be increased it is necessary that works be published in English.

References

1. Pavani, A. M. B. and Mazzeto, A. C. E. 2010. Examining Accesses by Country and Language, presented at ETD 2010 – International Symposium on Electronic Theses and Dissertations, June, Austin, TX, USA. Available https://conferences.tdl.org/index.php/utlibraries/etd2010/paper/viewFile/34/53 and http://www.maxwell.lambda.ele.puc-rio.br/Busca_etds.php?strSecao=resultado&nrSeq=16848@2.
2. Pavani, A. M. B. 2010. Examining Accesses by Country, Language and Area of Knowledge, presented at ETD 2011 – International Symposium on Electronic Theses and Dissertations, xx Sep, Cape Town, South Africa. Avalable http://dl.cs.uct.ac.za/conferences/etd2011/papers/etd2011_pavani.pdf and http://www.maxwell.lambda.ele.puc-rio.br/Busca_etds.php?strSecao=resultado&nrSeq=18638@2.
3. UNDP – United Nations Development Program 2010. 2010 Human Development Report. The Real Wealth of Nations: Pathways to Human Development, Available http://hdr.undp.org/en/reports/global/hdr2010/.
4. UNDP – United Nations Development Program, HDI HDI – Human Development Index, 1990. Available http://hdr.undp.org/en/statistics/indices/hdi/.