IFLA WLIC 2013
SINGAPORE

Submitted on: February 6, 2013

Virtual Libraries in Research Funding Agencies: an Open Source approach to disseminate Information to Faculty, Research Teams and Civil Society

Diego Ferreira Ucha
Virtual Library, São Paulo Research Foundation (FAPESP), São Paulo, Brazil.
E-mail address: diego@fapesp.br

Guilherme Giacchetto Moreira
Virtual Library, São Paulo Research Foundation (FAPESP), São Paulo, Brazil.
E-mail address: giacchetto@fapesp.br

Rosaly Favero Krzyzanowski
Virtual Library, São Paulo Research Foundation (FAPESP), São Paulo, Brazil.
E-mail address: rosalyfk@fapesp.br

Inês Maria de Morais Imperatriz
Virtual Library, São Paulo Research Foundation (FAPESP), São Paulo, Brazil.
E-mail address: immi@fapesp.br

Abstract:
Virtual Libraries lack of Open Source solutions when compared to the Digital Libraries context. The latter have a well-honed community to support these information systems such as DSpace, Greenstone and Fedora. As a solution for Virtual Libraries in Funding Agencies, São Paulo Research Foundation (FAPESP) has developed a Virtual Library to store and index the metadata for its funded Research Projects, Scholarships and Scientific Publications. This Virtual Library is able to contribute in addressing the needs of Civil Society, which can access referential information on the results achieved in funded research (taxes); of Academia since all the metadata for scientific projects, scholarships and publications are available in the World Wide Web freely and without limitations; and for the Funding Agency's staff which is able to analyse subject patterns in scientific research in a region and to assess the results for each ongoing/completed research project.

FAPESP Virtual Library has been developed since 2004, and continues to be upgraded and updated, with Open Source software. It was created on top of the Django Web Framework, which is written in Python, uses MySQL as a Relation Database Management System, Apache Solr as a search server and other Python/Django modules developed by the Open Source community, such as Haystack, which connects Django to Solr easily. FAPESP Virtual Library achieved more than 4 million page views in 2012 and its software is now in process to be shared with other Funding Agencies in Brazil, some of them have already informed on their interest.

Keywords: Virtual Libraries; Open Source Software; Research Funding Agencies; Research Projects; Scientific Publications.

1 INTRODUCTION

One of the definitions of Virtual Libraries (VL) is that of web information systems that provide access to a centralized database of resources (e.g. metadata), that is usually scattered on a network of systems (Marchiori, 1997). These resources are indexed, organized and available readily and economically.

On the other hand, Digital Libraries assemble rich digital collections (e.g. full text documents, manuscripts, high definition photos) (Zhang, 2010). It provides a similar organization and information retrieval capability when compared to virtual libraries.

The Digital Library context has a well-honed community to support its Open Source information systems, such as DSpace (Smith, Barton, Bass, Branschofsky, McClellan, Stuve, Tansley, Walker, 2003), Greenstone (Witten, Boddie, Bainbridge, McNab, 2000) and Fedora (Staples, Wayland, Payette, 2003). On the other hand, Virtual Libraries lack of Open Source solutions that adopt state of the art technology in information retrieval, storage, administration and so on.

As a solution for Virtual Libraries in Funding Agencies, São Paulo Research Foundation (FAPESP) has developed a public Virtual Library[1] to store and index the metadata for its funded Research Projects, Scholarships and Scientific Publications. This Virtual Library is able to contribute in addressing the needs of Civil Society, which can access referential information on the results achieved in funded research (taxes); of Academia since all the metadata for scientific projects, scholarships and publications are available in the World Wide Web freely and without limitations; and for the Funding Agency's staff which is able to analyse subject patterns in scientific research in a region and to assess the results for each ongoing/completed research project. This VL system is now in process to be shared with other Funding Agencies in Brazil, some of them have already informed on their interest.

The next sections will describe: i) how VLs can aid Research Funding Agencies in accomplishing the scientific information dissemination goal and, more specifically, in São Paulo Research Foundation (FAPESP) context; ii) the proposed Virtual Library System; iii) future works to be developed; and iv) the conclusion with results and recommendations.

2 VIRTUAL LIBRARIES (VLS) IN RESEARCH FUNDING AGENCIES (RFAS)

As stated before, VLs provide access to a centralized database of resources. In RFAs, VLs store metadata for each funded Research Project and/or Scholarship. In some cases, there are Funding Agencies (e.g. US National Science Foundation and Swiss National Science Foundation) that are also able to gather, store and display publicly the publications' metadata for each funded grant.

The way to gather these publications can be through manual or automatic means. The manual would consist, for example, of each researcher providing a list of scientific publications; while the automatic way would require a crawler that searches for patterns in the interested scientific publications' repositories/databases. The Swiss National Science Foundation addresses this task through the manual approach (Swiss National Science Foundation (SNSF), n.a.), while São Paulo Research Foundation (FAPESP) addresses this task through the automatic approach, which is detailed in section 3.2.

[1] FAPESP Virtual Library website: http://www.bv.fapesp.br/en/

Usually, the features available in RFAs VLs are: free and advanced search, search refinement, result reordering, results per page modification, download search results in text formats and a link to access each detailed search result. In this sense, RFAs VLs are a powerful tool to disseminate information, in the World Wide Web, regarding funded research, since each user is able to retrieve the exact information he needs. Since one of the main goals of each Research Funding Agency is to provide information to civil society about the results achieved, these VLs features help to accomplish the information dissemination goal.

2.1 VL IN SÃO PAULO RESEARCH FOUNDATION (FAPESP)

FAPESP Virtual Library has been developed since 2004, and continues to be upgraded and updated with open source software. It has achieved more than 4 million page views in 2012, which is a metric that is rising year after year.

The São Paulo Research Foundation (FAPESP) is one of the leading agencies that fund scientific research in Brazil, supporting research in all fields of knowledge, scientific exchange and the dissemination of science and technology. Its mission is to foster scientific research by awarding scholarships, fellowships and grants to investigators linked to higher education and research institutions in the State of São Paulo. It was initiated in 1962 and under the state constitution, 1% of all state taxes are appropriated to fund the Foundation.

Besides accomplishing the goal of providing public information access to the funded research projects and scholarship, FAPESP VL also assists in the following main tasks:

- Assessment of Research Programs' grants in a geographical and historical basis.
- Grant candidates evaluation.
- Internationalization assessment.
- Pattern identification in scientific publications resulted from funded grants.

We have identified that by providing public access to the research projects data, through the VL, the users on the World Wide Web end up using this data in interesting ways, such as i) by using some part of the research projects abstracts to answer questions in forums and Q&A tools; ii) a bibliographic reference in Wikipedia's article; iii) a Curriculum Vitae for funded researchers.

3 PROPOSED VIRTUAL LIBRARY SYSTEM

The proposed Virtual Library System has been developed to address the needs of a Research Funding Agency, more specifically to address the needs of São Paulo Research Foundation (FAPESP).

The adoption of Open Source software in order to build this System has proven to ease the effort necessary of the development team, since they could debug and, in some cases, tweak the code to their needs, for each Open Source software adopted (detailed in 3.1). This task is not so easily accomplished with Proprietary Software, which doesn't provide the possibility to analyze its internal source code.

3.1 MAIN OPEN SOURCE SOFTWARE SOLUTIONS ADOPTED

The proposed VL system was built on top of Django Web Framework[2], which is written in Python[3]. Both Python and Django have the philosophy of saving the developers' time in their daily work. For instance, Django comes with an automatic admin interface (Django Software Foundation, 2013) and also Python programs end up with fewer lines of code when compared to other programming languages (Norvig, n.a.). This way, the proposed VL system has an easier maintenance for the developers' team. This characteristic is important since systems tend to grow in features as the time passes by and yet the team will be able to work fast.

MySQL[4] was selected as the Relational Database Management System, which is a popular Open Source software with an active community (Oracle Corporation, 2013). Django has a built-in integration with MySQL, which makes the deployment easier.

Apache Solr[5] was selected as the search platform, in order to boost speed in search results and to provide the user with state of the art features in information retrieval, such as a spelling correction, stemming, faceting and filtering. The integration solution between Apache Solr and Django was Haystack[6] for the Django layer and PySolr[7] for the Python layer.

As a way to speed up web page delivery to the user and optimize the server performance, Memcached[8] is adopted as a memory object caching software. It also has an easy integration with Django, throughout its Cache Backend.

This whole system was deployed in a Linux server using Apache HTTP[9] Server through Web Server Gateway Interface (WSGI).

3.2 SYSTEM ARCHITECTURE

The figure 1 displays the diagram of the system architecture for the proposed Virtual Library. Each software in this diagram is detailed in section 3.1.

[2] Django website: https://www.djangoproject.com/
[3] Python website: http://www.python.org/
[4] MySQL website: http://www.mysql.com/
[5] Apache Solr website: http://lucene.apache.org/solr/
[6] Haystack website: http://haystacksearch.org/
[7] PySolr website: https://pypi.python.org/pypi/pysolr/
[8] Memcached website: http://memcached.org/
[9] Hypertext Transfer Protocol (HTTP)

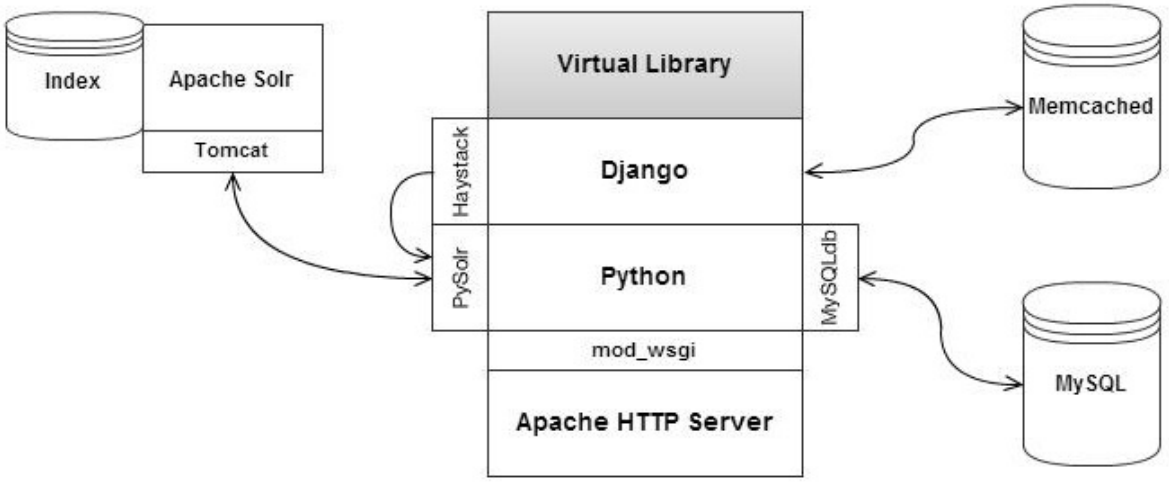

Figure 1 – Proposed Virtual Library System Architecture

The adopted version for each software are Apache HTTP Server 2.2, Python 2.6, Django 1.3, PySolr 2.0.15, MySQLdb 1.2.3, Haystack 1.2.7, Tomcat 6, Apache Solr 3.5, MySQL 5.1 and Memcached 1.43.

Some of the software solutions above can be replaced by other Open Source solutions. This is ideal for organizations that already have a well defined software infrastructure. The components that could be replaced are detailed below.

- Apache HTTP Server + mod_wsgi could be replaced by other HTTP Servers and WSGI Servers, such as Nginx (HTTP Server)[10] + Gunicorn (WSGI Server)[11].
- MySQL could be replaced by PostgreSQL[12] or SQLite[13], since both of them are supported in Django. For each one, there is a Python Database Binding that would replace MySQLdb. For PostgreSQL the binding is postgresql_psycopg2 and SQLite already has a built-in binding in Python 2.6.
- Memcached which stores its data in RAM Memory could be replaced by a database or a filesystem caching, both of them are available in Django's Cache Backend.
- Apache Solr + Tomcat + PySolr could be replaced by Haystack's supported search engines, which are ElasticSearch[14], Whoosh[15] and Xapian[16].

3.3 FEATURES

In the last paragraph of section 2, we have discussed about the common features available in RFAs VLs. Below are highlighted the specific features available in this proposed VL. The Open Source solutions described in chapter 3.1 and 3.2 were essential in the development of the features below.

[10] Nginx website: http://nginx.org/
[11] Gunicorn webste: http://gunicorn.org
[12] PostgreSQL website: http://www.postgresql.org/
[13] SQLite website: http://www.sqlite.org/
[14] ElasticSearch website: http://www.elasticsearch.org/
[15] Whoosh website: https://bitbucket.org/mchaput/whoosh/wiki/Home
[16] Xapian website: http://xapian.org/

i. Scientific publications gathering

We have developed an automated system that collects scientific publications funded by FAPESP and available in Web of Science (WoS). This process can be divided in two steps, as described below.

In the first step, it queries WoS for the entries of FAPESP, acknowledged by the authors in their publications, in the filter named "Funding Agency". For each range of result, it exports a BibTeX format file.

In the second step, the system parses all the BibTeX files and, if not yet available in the VL, imports the metadata to the VL database. An important metadata field in this process is the "Grant number". This field will be the one to create the relationship between the Grant and the Scientific Publication, i.e. it will be possible to identify the Scientific Publication as a result of a specific Grant.

This is one of the key processes in a RFA VL, since it will be able to publicly show to civil society and academia the results achieved by each funded Grant.

ii. Funded researchers' Curriculum gathering

The majority of Brazilian Research Institutions adopt a Federal Funding Agency solution for Web Curriculum Vitae called "Plataforma Lattes"[17], in which researchers are asked to register.

In order to provide more information about each funded researcher, the VL displays a link to their Lattes. To accomplish this, it was developed an automated system that queries Lattes for each researcher link. The collected links are then displayed in each researchers' individual web page in the VL.

iii. Individual pages for specific metadata fields

In the first version of the proposed VL, the search results for keywords which represented name of researchers, grant's knowledge areas or research subjects were displayed to the user as a standard search result page.

In 2011 and 2012, it was developed the individual pages for these specific metadata fields. These individual pages summarize the content available in the VL in a more comprehensive way than a list of search results.

This solution has proven to be a great way to improve information access. As an example, the researchers' individual pages already represent more than 21% of the overall access. The researchers' pages, for instance, contain the researcher's short résumé, his photo, a list of all funded grants where he participated, a list of the most frequent collaborators in the funded grants, links to Thomson Reuters's ResearcherID and Google My Citations.

iv. Heat map of funded grants per city of the State of São Paulo

In each individual page for specific metadata fields, as described in topic iii, a heat map showing the State of São Paulo is displayed with the identification of grants concentration in a municipality basis.

[17] Plataforma Lattes website: http://lattes.cnpq.br/

This feature uses Google Maps API[18] to render the map and, in an offline automated procedure, it collects the latitude and longitude for each State of São Paulo municipality that has a Research Institution that hosted a Grant.

The map could be centered in other regions of the globe, being only necessary to provide the latitude, longitude and weight (e.g. quantity of funded grants) for each highlighted point.

v. Historical view of grants concession throughout the years

This feature shows a dynamic chart, by using Google Charts API, where it displays the number of grants awarded per year. This feature makes it easier to understand historical patterns when assessing, for instance, Special Research Programs or specific Research Subjects. A type of assessment would be to evaluate the evolution or decrease in an historical window.

vi. Visibility boost in search engines

It was first introduced in 2009 an optimization for search engines of all the pages of the proposed VL. The optimization is a technique called Search Engine Optimization (SEO) (Grappone & Couzin, 2010), which focus on adapting the web pages to be better indexed by search engines' crawler.

This optimization has boosted the visibility of the VL. The access increase when comparing 2009 with 2008 was of 896% and comparing 2012 with 2008 it was 2,641%. Each year has registered an expressive increase in absolute numbers.

vii. Internationalization[19]

As the main feature to disseminate information to a wider range of users on the Web, the proposed VL system is capable of delivering metadata information in multiple languages. In FAPESP VL, the adopted languages are Brazilian Portuguese and English.

One of the supported built-in features in Django is the internationalization process that makes the translation job easier. All strings that are explicitly marked to be translated will be copied into a unique file, in a language basis, in order to be translated by the translator team. Once translated, the developer must compile the files, in order to enable Django to automatically replace the marked strings in each web page.

viii. Grant pages' access statistics available to the RFA's staff

The RFA's staff is able to assess the Web access statistics of each Grant available through the VL. This feature integrates with Google Analytics API to gather the Web access statistics of the visualized grant's page and, throughout Google Charts API it displays the data with dynamic charts.

[18] API stands for Application Programming Interface
[19] Internationalization as a mechanism to multiple language support in a system

This feature is essential to ease the assessment of each page's trends in information access in an easy way to a RFA's staff. It also eliminates the need of training each individual in web analytics tools.

ix. VL staff's administration area to create, edit and remove content

The librarian staff is able to create, edit and remove content from the VL using an administrative area, by authenticating their credentials through a login page.

This administrative area was created using the built-in Django's features. By modeling the system, Django is able to generate an automatic admin interface. It also enables the developers to customize this admin interface as needed, in a project basis. This built-in feature saved a considerable amount of the developers' time, since they didn't have to develop great part of the admin functionalities.

One example of a developed feature for the VL is the Librarians Production Reports, in which it is able to assess the day-by-day work of the librarian staff. It also saves the staff time since they won't have to keep track of the work done by them in a daily basis.

x. Sending email alerts to subscribers

In any search result, the user is able to register his email to receive the new grants entries in the VL. The new entries will only be emailed if they correspond to the search result's keywords, provided by the user.

Once the user inputs its interest by registering his email in a specific field on the interface, the system registers this data, along with the keywords inputted by the users, as well as the selected refinements, and stores these data on the database. Once a week, an automated system queries this database and checks for the new grant entries, in the VL, since the last email alert issued for each user. The new grants are selected, in a user basis, and the process finishes with the personalized email sending.

4 FUTURE WORK

An important step for the proposed RFA VL System is to be shared freely among other RFAs. This process is about to start with some RFAs in Brazil, since FAPESP is working to settle cooperation agreements with these RFAs.

The future work in this system will be focused in implementing more graphical summarizations of data to ease the decision making process of a RFA's staff and to ease the information access to civil society and academia.

5 CONCLUSION

Although the Digital Library context has a well-honed community to support its Open Source information systems, the Virtual Library context lacks of Open Source solutions that adopt state of the art technology. A solution to Research Funding Agencies (RFA) Virtual Libraries (VL) would be the proposed VL in this work, that assembles Open Source solutions that have a well-honed community of developers in order to deliver a high impact RFA VL to the civil society, academia and to its staff.

6 REFERENCES

Django Software Foundation. (2013). *The Django admin site.* Retrieved March 13, 2013, from Django: https://docs.djangoproject.com/en/1.5/ref/contrib/admin/

Grappone, J., & Couzin, G. (2010). *Search Engine Optimization (SEO): An Hour a Day.* Indianapolis: Wiley Publishing.

Marchiori, P. Z. (1997). "Ciberteca" ou biblioteca virtual: uma perspectiva de gerenciamento de recursos de informação. *Ciência da Informação, 26*(2).

Norvig, P. (n.d.). *How to Write a Spelling Corrector.* Retrieved March 13, 2013, from Peter@Norvig.com: http://norvig.com/spell-correct.html

Oracle Corporation. (2013). *Download MySQL Community Server.* Retrieved March 13, 2013, from MySQL: http://dev.mysql.com/downloads/mysql/

Smith, M., Barton, M., Bass, M., Branschofsky, M., McClellan, G., Stuve, D., et al. (2003). DSpace: An Open Source Dynamic Digital Repository. *D-Lib Magazine, 9*(1).

Staples, T., Wayland, R., & Payette, S. (2003). The Fedora Project: An Open-source Digital Object Repository Management System. *D-Lib Magazine, 9*(4).

Swiss National Science Foundation (SNSF). (n.d.). *Output of research.* Retrieved March 15, 2013, from Swiss National Science Foundation (SNSF): http://www.snf.ch/E/current/Dossiers/Pages/output-of-research.aspx

Witten, I. H., Boddie, S. J., Bainbridge, D., & McNab, R. J. (2000). Greenstone: a comprehensive open-source digital library software system. *Proceedings of the fifth ACM conference on Digital libraries* (pp. 113-121). New York: ACM.

Zhang, Y. (2010). Developing a Holistic Model for Digital Library Evaluation. *Journal of the American Society for Information Science and Technology, 61*(1), 88-110.