REPOX: Obtenção e agregação de dados para Indexação em Bibliotecas Digitais em Rede

Tiago João de Sousa Marques

Instituto Superior Técnico
Av. Rovisco Pais, 1049-1001 Lisboa, Portugal
tiago.marques@ist.utl.pt

Abstract. *With the advent of the Digital Libraries and Archives came the need to share with the community information that each of these entities has. This reality forced the creation of solutions that enable the dissemination of information present in each of these entities, so that it could be searched quickly and easily. However, due to the growth of new technologies, new goals were set by the entities listed above, passing by the possibility of researching information in the texts of the documents. Thus, this thesis will be focused on the demand for existing technologies and to obtain data synchronization and how these can be applied to solve facilitate the collection of new data content, for future indexation in search engine.*

Keywords: Digital Libraries, metadata, REPOX, content harvest (fulltext).

1 Introduction

The emergence of digital libraries and archives generated a great interest in sharing descriptive information in international projects such as Europeana[1], TEL[2] e EuDML[3]. These entities have the objective of harvesting this descriptive information in order to share it, making it easier to search. In these scenarios, the organizations willing to share their data often face problems doing it, due to their systems not support the commonly required OAI-PMH[1] protocol. This technology is supported by commercial and open-source solutions, but the required investment from the first not always possible and the lack of technical expertise for the local customization of

[1] Europeana –The European Digital Library -http://dev.europeana.eu/
[2] TEL - The European Library- http://www.theeuropeanlibrary.org
[3] EuDML– European Digital Mathematics Library- http://www.eudml.eu/

the open-source solutions are difficult barriers that the projects have to face. Also the emerging of new data transferring scenarios, such as the harvesting of content described by the data sets, will require the support of more sophisticated harvesting and aggregations processes.

In order to facilitate the sharing of data between digital libraries and other projects that promote the harvesting and aggregation of that data, an open-source platform with the name REPOX [3] was created. This framework was designed to address the problems of harvesting and aggregation of metadata that the data providers seek. It as also the ability to transform and publish the metadata, which is crucial to the service providers. This framework also intended to support a fast start process with a very easy installation and configuration that required little technical knowledge. However the actual versions of this framework only provide metadata harvesting solutions, which is not enough for the rising challenge of harvesting storing and updating the full-text that is described in the metadata. So with this in mind a new solution must be designed in order to give to the REPOX capability not only harvest the metadata and its contents but also to give the ability to publish and share that same information, keeping the principals of the framework intact.

Overall, this search seeks to develop a solution that promote the harvesting of content referenced by metadata, keeping it actual and have the ability to share it and its actualizations in the most efficient way, promoting the interoperability between service providers and data providers.

2 Related Work

Considering the objectives defined in the context of the digital libraries and the data aggregation projects, it is possible to define which themes should be studied in order to try to find a solution passable of be integrated in the REPOX project. Taking that in to consideration the themes that were approached are: Data and File synchronization and Data-sharing Middleware. The principal solutions studied where SyncML[4], Rsync[5], Unison[6], Semantic-Chunks[8], IceCube[7] and the Xmiddle[9].

In conclusion is possible to see that all the solutions studied have the capability of replicating and synchronizing files and folder between different environments and devices ensuring its consistency, but in none of the cases we were able to find solution for the following objectives:

- harvest of contents from references stored in metadata
- synchronization of updates made to the contents harvested from the referenced objects.

The principal reason that excludes this technologies from being a valid solution to the problem that is being presented is that their thought to work in communication with the origin of the files from where they can propagate or receive updates to the data that they have stored, as in opposition to what happens in the scenarios presented by the digital libraries, where the data is referenced in the metadata, and the only way to get the object is through direct download of that instance, making it almost

impossible to know when a new version was stored. This limitation makes the use of any one of studied technologies unfeasible due to their inability to face the presented challenges.

So the only solution, based on the conclusions stated before, is to develop a technology aggregated to the REPOX, in order to benefit from its specificities to resolve the problems of harvesting and storing metadata, that makes the download of the referenced objects that are stored in the data harvested by framework and guarantees the synchronization of the downloaded data with the service and data providers.

3 Implementation and Solution

In order to handle the objectives defined along the document, we have designed a simple solution that is constituted by a set of elements that will bring the flexibility and the scalability to ensure that independently of form of the metadata and the location of the descriptive information it will be always possible to harvest the objects that are implicit in the metadata.

The figure 1 represents the defined architecture for the produced solution.

Has it can be seen in the figure mentioned previously the solution is divided various elements, that will serve the purpose of harvesting and synchronizing objects referenced in descriptive data.

The main component in the architecture is the Harvester because it contains the principal functions implemented. These are the Harvest of data, the synchronization of the data and the re-harvest of data in case there was something wrong with the link to the object or with object himself.

The process of harvesting is defined by the configuration of the data set that will provide the metadata or the configuration that is implicit in metadata that was previously harvest by other forms than the protocol OAI-PMH. For each form there are three mandatory pieces of information to realize the harvesting process. These are the name of the set of the metadata, the schema of the metadata and the xpath necessary to travel the descriptive data to find the wanted objects to download. With this information it will be possible to create the information structures necessary to perform the harvesting of the data and generate logs to report the events that happened during the process.

The process of synchronization and of the re-harvest of the errors e based on the result of the harvesting process. If the harvest has ended successfully it will be possible to verify if there are any updates made to the objects, ensuring the synchronization of the same, using a conjunction of three techniques: verification if any of the records obtained by OAI-PMH has updates, verification of the last modified date of the object, if that field is available to analyze and finally the verification by checksum of a new downloaded file, which in the case of download of a new file the older is stored and new one takes the place of the other one. In the case of the harvesting process has ended with errors it will only be possible to try to

correct the errors by running the harvesting process through the failed files, insuring that it will only be harvested the necessary information to terminate the harvesting of the set.

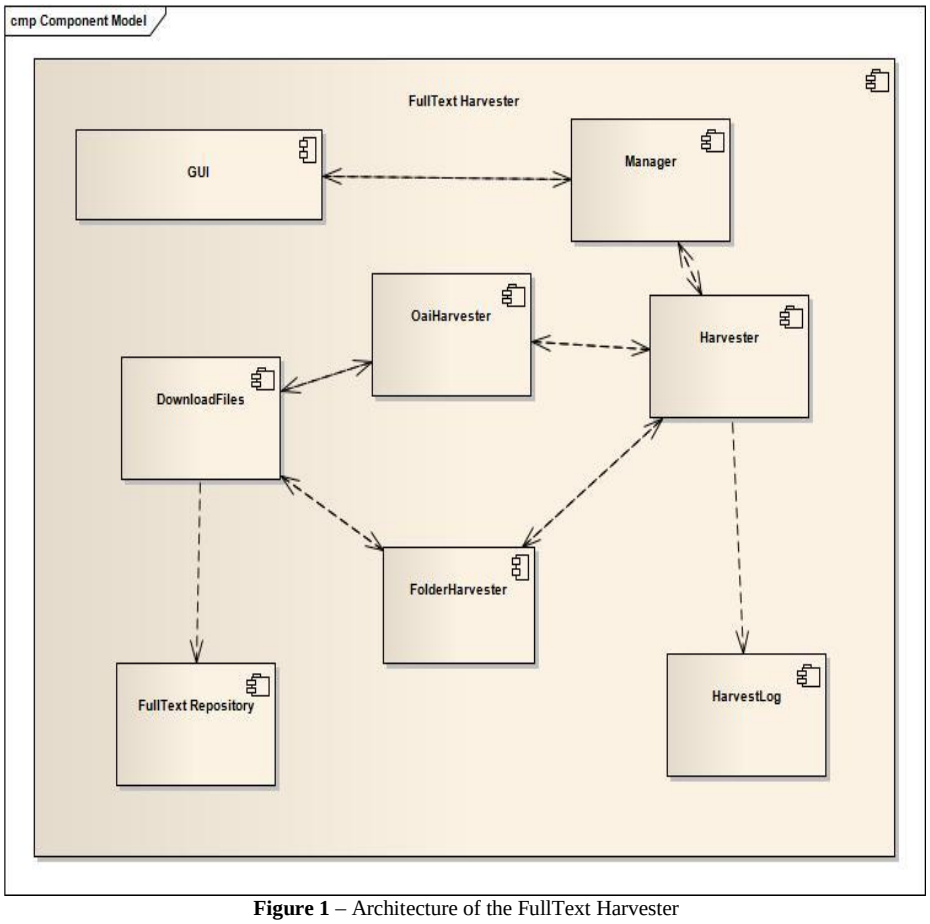

Figure 1 – Architecture of the FullText Harvester

4 Practical Results

In order to validate the developed solution, the REPOX framework was used, to have access to different scenarios such as the Europena. TEL, EuDML ou SHAMAN. The main scenario used to test the solution was the EuDML due to is best rate of availability of the contents to download and has a different set of formats that allow us to capability of the solution.

We evaluated the principal scenarios regarding the harvesting, synchronization a re-harvesting of objects. Has an example of a harvesting process (**Figure 2**) we decide to harvest multiple data sets to show clearly the quantity of information that is downloaded and generated information from the execution of each data set.

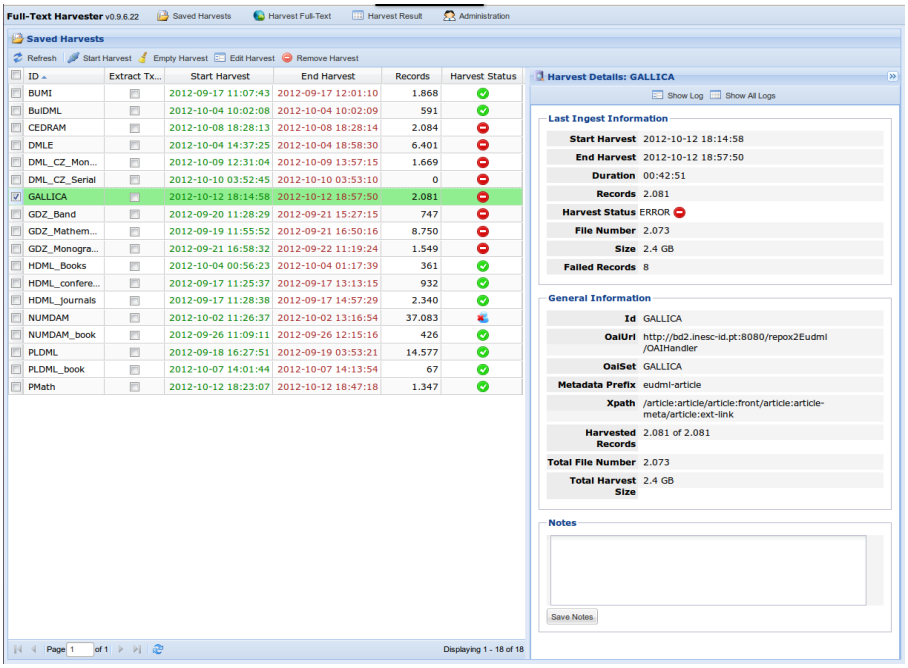

Figure 2 – Complete harvest with multiple data sets

In the data sets that being are displayed it possible to see the state of each data set, and the amount of records that where harvested which by comparing the number of records that are stored in the REPOX it is possible to verify that they match allowing us to confirm the successful creation of a solution that has the ability to harvest and maintain the quality of that referenced objects downloaded.

5 Summary and Future Work

With the appearing of the digital libraries and archives, new paradigms of harvesting and aggregation of metadata have emerged. These necessities created REPOX, a framework designed to adequately deal with this challenges and promote the interoperability between service providers and data providers. With the evolution of technologies the service providers decided that the metadata will have references to the objects that originated the metadata. This has created a new an interesting challenge, to harvest the referenced objects in the descriptive data, challenge that the REPOX technology was not up to, but needed to be capable to respond.

So in general terms we have created a simple but powerful solution that manages the harvesting of the referenced objects in the descriptive data, which gives the possibility of monitoring the evolution of the harvesting process and to check for updates from the referenced objects, keeping the information as fresh as possible. This was validated in real scenarios such as EuDML, which allows confirming that the solution is viable and it is a valid approach to solve the problems presented along the document.

In the future, the interoperability using the harvested referenced objects should be introduced, in order to facilitate the share of information, and streamline the process of updates of that information, reducing the number of total downloaded files to the max. Other work that can be developed should be the possibility to schedule harvests and/or updates of files, allowing the users to program in advance the work that should be done, in a case of unavailability of the referenced objects at the time of the download order.

6 References

[1]Carl Lagoze and Herbert Van de Sompel. The open archives initiative: building a low-barrier interoperability framework. In JCDL '01: Proceedings of the 1st ACM/IEEE-CS joint conference on Digital libraries, pages 54–62, New York, NY, USA, 2001. ACM.

[2]Van de Sompel, Lagoze. D-Lib Magazine February. 2000, Vol. 6 Number 2.Consultado em: http://www.dlib.org/dlib/february00/vandesompel-oai/02vandesompel-oai.html

[3]Freire, Manguinhas, Borbinha, REPOX: Uma infra-estrutura XML para a PORBASE, Lisboa

[4]SyncML White Paper:Building an Industry-Wide Mobile Data Synchronization Protocol

[5]Andrew *Tridgell and Paul Mackerras. The rsync algorithm. The Australian National University*

[6]David *Rasch and Randal Burns. In-Place Rsync. File Synchronization for Mobile and Wireless Devices. Department of Computer Science Johns Hopkins University*

[7*]A.-M. Kermarrec, A. Rowstron, M. Shapiro, and P. Druschel. The icecube approach to the recon-ciliation of divergent replicas.*

[8*]L. Veiga and P. Ferreira. Semantic-Chunks a middleware for ubiquitous cooperative work. In Proceedings of the 4th workshop on Reflective and adaptive middleware systems, page 6. ACM, 2005.*

[9]S. *Zachariadis, L. Capra, C. Mascolo, and W. Emmerich. XMIDDLE: A Data-Sharing Middleware for Mobile Computing. 2002*