Proposal of a Data Warehouse for Scholarly Institutions built on Institutional Repositories

Pablo C. de Albuquerque [1,2] [0000-0001-5277-1665] , Gonzalo L. Villarreal [1,2] [0000-0002-3602-8211] , Marisa R. De Giusti [1,2] [0000-0003-2422-6322]

[1] PREBI-SEDICI, Universidad Nacional de La Plata. La Plata, Argentina.
[2] CESGI, Comisión de Investigaciones Científicas. La Plata, Argentina.
pablo@sedici.unlp.edu.ar, gonzalo@prebi.unlp.edu.ar, marisa.degiusti@sedici.unlp.edu.ar

Abstract. A Data Warehouse (DW) is a tool that integrates and unifies information from multiple data sources and is used to assist decision making. In academic institutions, a Data Warehouse oriented to scientific and academic intellectual production could provide valuable information to understand, optimize and promote the processes involved in intellectual production. This work proposes to use the data sources that conform the Institutional Repositories to start developing a DW.

Keywords: Data Warehouse, Institutional Repositories, Business Intelligence

1 General Data Warehouse Concepts

A Data Warehouse (DW) is a tool that integrates and unifies information from different data sources of an organization, and serves and is useful for decision making. Data sources are usually heterogeneous, both from the point of view of the technological support (e.g., relational databases, NoSQL databases, spreadsheets, text files, etc.) and also from the point of view of the purpose of each source (for example transactional management systems, monitoring services, server access logs, etc.). The integration of these data sources into the DW is done by retrieving or extracting data from those sources, which are then transformed and finally integrated into a centralized database; this process is known as ETL: Extract-Transform-Load.

One of the DW design premises is to keep a simplified data model, requiring simple queries to retrieve useful information. This simplicity ease the integration of the DW with different Business Intelligence (BI) and/or reporting systems, such as Power BI, Google DataStudio or even MS Excel, and also promotes the exploitation of the data by users who have elementary concepts but are not necessarily database experts.

The volume of data in a DW usually grows rapidly and in many cases at an accelerated rate, reaching the order of GB, TB or even EB in shor time. Despite its size, the DW must be able to execute queries and return results in optimal response times. To achieve these requirements, many actions linked to the optimization of the underlying tools (server, database engine, network, etc.) must be combined with the design of the

DW database itself, usually based on a denormalized star model, built on facts and dimensions associated with the facts. [1]

2 Data Warehouse in Scholarly Institutions

In scholarly institutions, a Data Warehouse focused on scientific and academic outputs could provide valuable information to understand, optimize and promote the processes involved in their production: what type of resources are produced, what are the areas of research, who conducts the research, where they are produced from (research centers, departments, editorial teams), when the different resources are generated, what mechanisms are used to produce or publish the resources, and how they are used both internally (research projects, working groups, theses, etc.) and externally (citations, visualizations, downloads, mentions, etc.). [2]

Like any organization, most scholarly institutions have a wide diversity of systems that manage, host and publish different resources produced by the institution. Like expected, each system organizes and manages its data based on its needs and the availability of technological resources at the time of the development. That is why the diversity of data sources that make up the ecosystem of technologies around an institution can make certain tasks more difficult such as assisting in decision making, since it is necessary to integrate these sources in a single place. Some of the typical systems used in scholarly institutions include institutional repositories, current research information systems (CRIS), journal portals, conference portals, digital book portals, among others. It is important to be clear that not only the information directly associated with the function of each system is important (for example, books and their authors in a Books Portal), but also much information generated by the system itself: users' access, server logs or even security reports linked to each system.

Following is a description of some of these systems, with emphasis on what data they manage, how reliable they are, and what processes should be implemented to integrate these data into the Data Warehouse:

- Institutional Repository (IR):
    - Advantages: data are already standardized through the use of multiple controlled vocabularies, reviewed by staff dedicated to ensuring compliance with repository policies, and adoption of guidelines that allow integration into repository networks. The organization into collections and communities provides valuable information. The use of persistent identifiers makes it possible to identify a resource univocally on the web, facilitating interoperability with other systems. As mentioned above, a repository can be part of repository networks that provide services and increase the visibility of the scholarly production. Repositories can also participate in different agreements with other institutions, which gives access to standardized resources that, after being reviewed, can be incorporated into a particular collection. The adoption by the institution's users is also important since many are already using these services. [3]

    - Disadvantages: authors do not always deposit their production in the IR which could generate a partial view. Besides, many repositories include "less interesting" resources such as learning object or internal lecture notes.
- CRIS System
    - Advantages: the amount of data these systems usually store tends to be very complete, since these systems are generally used for institutional evaluation tasks and therefore the different actors of the institution must ensure that the results of their work are there for the evaluators.
    - Disadvantages: the information is uploaded by the author who is generally not skilled in describing the resources that are submitted, and there are usually no instances of review of this data. Many data will be repeated among authors, in part because no identifiers are used to create relationships between resources and people (authors, editors, etc.). In general, the data are not standardized.
- Books and Journals Portals
    - Advantages: Similar to the IR, these portals have reliable data, uploaded by the authors and in this case corrected by the different editorial teams. Their organizational structure is usually simple: numbers, volumes, articles, in the case of journals; academic units, thematic areas in the case of book portals. These systems use persistent identifiers, which improve interoperability. In many cases they provide data on how these resources were generated.
    - Disadvantages: not all editorial teams will necessarily have the same policies and quality in their metadata, nor will they follow the same workflows. It should also be noted that systems based on standardized metadata schemas are not always used.

While there may be many other systems in these institutions, it seems clear that the IR is a great candidate to begin the development of a DW: the volume of information that an IR can handle, the reliability of the stored data, the available interoperability tools, and the existing services and infrastructure provide an interesting starting point.

3 Users and roles

As mentioned above, around a repository there are actors with different responsibilities and needs that periodically require access to information to assist them in their decision making. The IR provides data that allows them to prepare reports, analyses and dashboards that reflect the reality of a part of the repository at a given time. Some examples of data requirements to an IR are repository managers may need to know the impact factor that was generated by the import of a new collection into the repository; technical staff may want to know how many requests are served by the web server in the last month and of that total, which is the flow of malicious bots identified by a third-party service and then, based on this information, make decisions that allow filtering and maintenance of the infrastructure; the administrative staff working in the

repository may want to know the status of the resources imported in the last year, in order to know if they should perform revision tasks on them; the authorities of the institution may need to see the growth in the number of items by typology in the last semester, by institution, department or academic unit; the authors of the resources stored in the repository may want to know from where the resources they have participated in have been accessed; visitors who search and download resources from the repository may want to see where the research lines of the different academic units are progressing. Although these are just a few examples, it can be seen that some IRs are already fulfilling the functions of a DW. However, as they only have their own internal data, they do not provide an overall picture of the entire institution.

4 Putting it all together

In this work, we have reviewed some of the functions and requirements typically served by an IR, with emphasis on the information requests and reports that may be solicited periodically. As we have mentioned, each system or data source structures its information to respond to its own needs, so some data may not be directly available and may require a special process to be inferred or calculated.

While many of the above tasks can be automated and scheduled, it is important to keep in mind that they always involve data from the repository itself, but it is often necessary to combine data from other data sources to get a complete picture.

To solve this, it may be necessary to use other sources, so it is no longer sufficient to define tasks that process information from one source to infer other data, but rather to define processes that unify and standardize various sources in one place.

A Data Warehouse would solve these problems, gathering in one place the necessary information to have a broader view of the academic situation of an institution, simplifying the tasks of data integration and normalization, with the aim of answering queries to users, such as institutional authorities, technical and administrative staff and the general public, in order to assist them in their decision making. The development of a DW implies a great effort, both on the part of the team responsible for its design, implementation and maintenance, as well as on the part of those responsible for the different areas of the institution whose data must be periodically integrated. For this reason, the success of such a project requires the commitment of the entire institution, from the highest authorities to the technical staff responsible for managing each database. However, the potential for obtaining useful, quality and instantaneous information from this kind of tool suggests that perhaps academic institutions should seriously consider investing resources in its implementation.

References

1. Kimball, Ralph, y Margy Ross. The Data Warehouse Toolkit: The Definitive Guide to Dimensional Modeling. John Wiley & Sons, 2013.
2. Rójas-Muñóz, C., & Saquicela, V. (2017). Sistema de ayuda a la decisión basado en un data warehouse. Maskana, 8, 175–187. Recuperado a partir de https://publicaciones.ucuenca.edu.ec/ojs/index.php/maskana/article/view/1461

3. De Giusti, Marisa Raquel. «Los nuevos roles del repositorio institucional». Visión Conjunta 10, n.º 18 (agosto de 2018). http://sedici.unlp.edu.ar/handle/10915/72087.