Data integration systems and bibliometrics

Samiya Tamtam[1*], *Ahmed* Laguidi[2, 3], and *Abderrafiaa* Elkalay[1]

[1]Laboratory Computer Science & Smart Systems (C3S) ESTC, Hassan II University Casablanca,Morocco.
[2]Laboratory of Networks, Computer Science, Telecommunication, Multimedia (RITM) ESTC, Hassan II University Casablanca, Morocco.
[3]MAEGE, Faculty of law, Economic and Social Sciences Ain Sbaâ, Casablanca, Morocco.

Abstract. The expansion and multiplication of information sources and their heterogeneity on the Web are one of the main difficulties encountered by users today. As a result, the integration of information is essential. It enables data from several sources to be brought together and consolidated. A concrete case of this diversity of information concerns the use of bibliometrics to analyze the literature indexed in bibliographic databases. Indeed, huge amounts of information are generated every day in scientific databases, and the proper analysis and decoding of this data are essential to uncover patterns of collaboration, emerging trends, research constituents, etc. These data would provide a solid basis for the development of new research projects. This information would constitute a solid basis for initiating a strategic and economic intelligence activity....

1 Introduction

Over the past decades, the exponential growth of information and communication technologies (ICT) has facilitated access to information and its production. Exploring these heterogeneous scientific data sources is now a primary requirement. Thus, it is important to develop systems that allow us to access these information sources, called data integration systems (DIS). Combining heterogeneous data sources and querying them via a single query interface is a difficult task.

Our work, therefore, aims at proposing a new architecture. To do so, we use data integration systems to perform bibliometrics. The objective is to process scientific questions from bibliographic databases such as Scopus, Web of science, or others. This process allows the

* Corresponding author: samiya.tamtam@gmail.com

scientific community and bibliometric practitioners to have data analysis to measure the scientific production and trend to make good decisions.

For this purpose, we conducted a literature review to identify previous works that have addressed the issue of integration systems in the context of bibliometrics.This study is based on the bibliographic databases Scopus and Web Of Science, limiting the search period to the last ten years.The query consists mainly of keywords specific to this theme, in English: "data integration", "hybrid integration system", "information integration", "mediator architecture", "database mediation", "information integration system personalization", "hybrid mediator", "hybrid integration", "Heterogeneous Data" and "bibliometric".

Based on this review of the various theoretical works, we can list the documents related to integrated systems and bibliometrics.

2 Data integration systems

2.1. Definition

Data integration (DI) refers to the combination of data from multiple sources (different databases, files, and even different formats) into a single, unified, comprehensive view. It allows for the integration of all types of data, considering their growth, volume, and varying formats. According to Lenzerini [1], the concept of DI raises the problem of combining different sources to provide users with a unified view of data.

At the architecture level of a DIS, there is the global schema. It provides a unified view of local sources and is also used to support queries. The architecture is illustrated in the following figure:

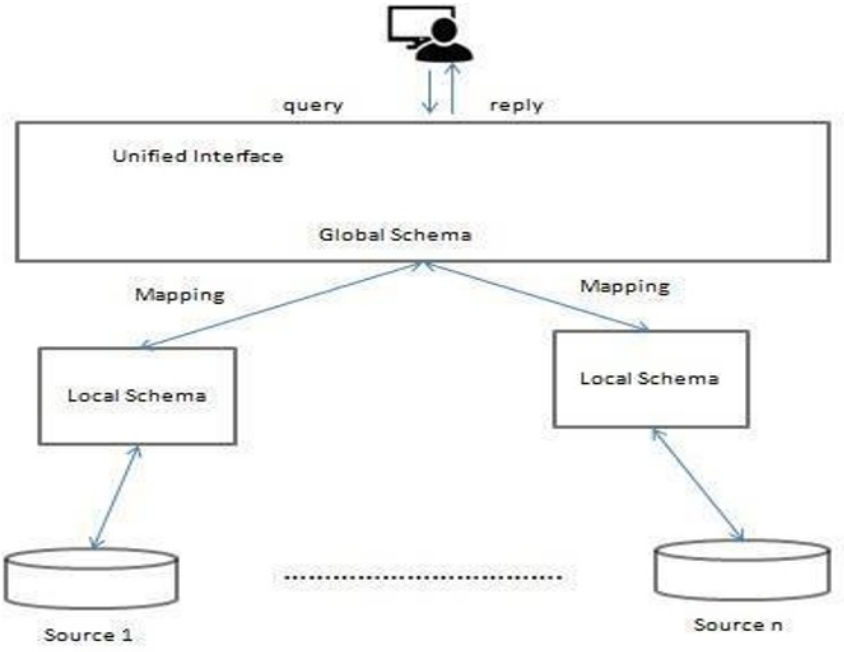

Fig. 1. General architecture of a data integration system [2]

2.2. Approaches to integrating heterogeneous data sources

Two approaches to integrating heterogeneous data exist depending on where the data are integrated [3]. Under the virtual approach, the data remains at the source.

Under the materialized approach, data is replicated in the data warehouse. Thus, a system based on a virtual approach is called a mediator system, while a system based on a materialized approach is called a data warehouse system [4].

2.2.1 Mediator Approach (MA)

It is "an approach providing an intermediary tool between users or applications on the one hand, and an autonomous, heterogeneous, distributed and evolving set of information sources on the other hand, this tool offers transparent access to the source via a single interface and query language" [5].

This virtual approach gives the illusion that users query a single homogeneous system when consulting distributed, autonomous and heterogeneous sources [6]. According to Wiederhold [2] the mediator refers to a software layer exploiting knowledge. Users, therefore, access distributed and heterogeneous sources transparently.

The MA thus provides a unified query interface from more meaningful queries, using the vocabulary of the global schema [2]. The latter virtually represents the data using abstract views that specify the content of the sources [2]. The architecture is illustrated in the following figure:

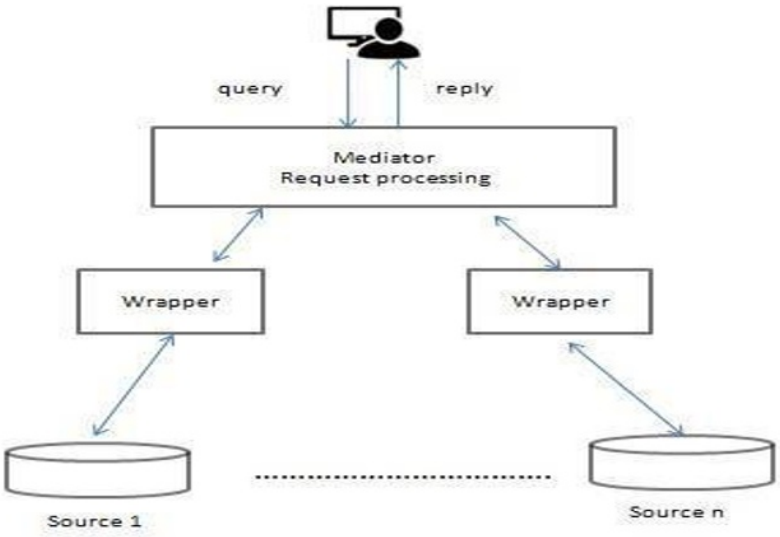

Fig .2. Architecture of the mediator

2.2.2 Data warehouse approach (DWA)

Unlike the MA, which queries data from its original source, DWA aims to establish a single database: the data warehouse from the data sources. The query is done through queries and the terminology used refers to the global Warehouse schema [7].

The illustration below represents the architecture of a data warehouse integration system.

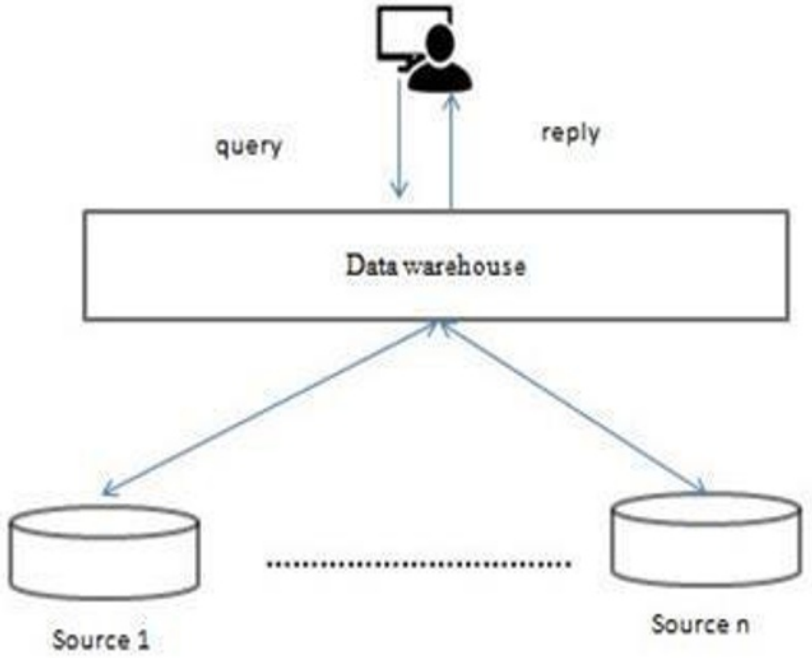

Fig .3. Data warehouse system architecture

2.2.3 Hybrid approach

A hybrid approach combines the MA for external sources and DWA for querying their data. The hybrid approach is defined as: "A system in which some data is queried on demand as in the virtual approach, while other data is retrieved, filtered, and stored in a local database" [8].

It corresponds to "systems in which some attributes are materialized and others are virtualized" [9].

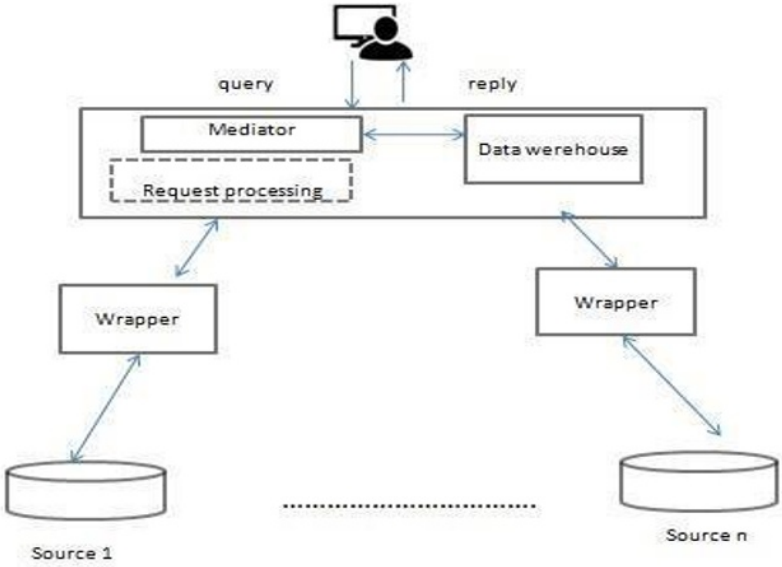

Fig .4. Hybrid integration system architecture

3 Conclusion and Perspectives

Bibliometric is the use of statistical methods to to measure the impact and influence of academic research it is a quantitative and qualitative analysis of data from publications indexed in scientific or technological databases and data from authors and institutions [10].

Bibliometric analysis is used to explore a specific area in the literature review to uncover collaborations, emerging trends, and constituents of research [11].

Bibliometrics can help answer important questions such as [12]:

- What is the annual scientific publication growth?
- Which authors are the most productive?
- Which collaborations are the most successful and produce the most influential and recognized work?
- Which journal do scientists mostly publish their articles?
- Who are the most cited scientists?

The application of DIS and bibliometrics can be used together to provide a more general view and analysis of the scientific literature in a particular field. For example, a DIS can use bibliographic data to create a database of scientific publications, which can be used to perform bibliometric analyses.

By integrating data from several databases; a researcher has a more complete understanding of the publications and authors in a particular field, as well as the relationships between them. It also provides visibility into trends in research topics and the impact of different researchers and institutions. Additionally, bibliometrics informs decisions about research funding, journal publication, and academic promotion.

Our future work focuses on proposing an integration architecture for heterogeneous data sources. The aim is to offer a solution to the various problems related to the heterogeneity of the sources (the exploration of bibliometric databases). The proposed solution is based on the MA. To validate and finalize the proposed architecture, we consider future works:

- Adaptedqueries
- Optimizedstorage
- Treatment and elimination of duplicates
- Preparation of data prior to exploration
- Processing to obtain homogeneous data from different sources
- Decoding of the processed data
- Data processing to create a prototype of a decisionmaking tool based on bibliometric analysis.

Based on the above, a combination of the mediator and ETL approach is very promising. The suggested model can solve many existing problems; however, practical validation of this model is important.

References

1. Lenzerini, «'Data integration: A theoretical perspective', In Proceedings,» In Proceedings of the twenty-first ACM SIGMODSIGACT-SIGART symposium on Principles of database systems, pp. 233-246, ACM., (2002).
2. W. G, «Wiederhold G. 'Mediators in the architecture of future information systems',,» *IEEE computers, Vol. 25, No. 3, pp. 38-49., (*1992).
3. C. e. S. S. (. Parent, «'Integration de bases de données: Panorama des problèmes et des approches', Ingéniérie des Systèmes d'information, Vol. 4, No. LBD-ARTICLE-1996-002, pp.333-359.,» (1996).
4. A. C. D. D. G. G. e. L. M. Calì, «'Data integration under integrity constraints', In Seminal Contributions to Information Systems Engineering, pp. 335-352, Springer Berlin Heidelberg,» (2013).
5. A. Zellou., «Contribution à la réécriture LAV dans le contexte de WASSIT, vers un Framework d'intégration de ressources ",» (2008).
6. M. G. M. M. e. H. M. S. Sellami, «'Secure data integration: a formal concept analysis based approach', In International Conference on Database and Expert Systems Applications, pp. 326-333, Springer International Publishing.,» (2014).
7. W. e. C. K. Inmon, «'RDB/VMS: Developing the Data Warehouse', Livre. Boston. QED Pub Group.,» (1993).
8. J. Widom, «, Integrating Heterogeneous Databases: Lazy or Eager?,» *ACM Computing Surveys 28A(4), (*1996).

9. G. Z. R. Hull, «" A Framework for supporting Data Integration Using the Materialized and VirtualApproaches",» (1996).
10. H. Rostaing, «La bibliometrie et ses techniques,» *https://hal.archivesouvertes.fr/hal - 01579948*, (2022).
11. S. K. D. M. N. P. W. L. N. Donthu, « How to conduct a bibliometric analysis: an overview and guidelines, ,» https://doi.org/10.1016/j.jbusres.2021.04.070., (2021).
12. H. Darvish, «Bibliometric Analysis using Bibliometrix an R Package,» *Journal of Scientometric Research 8(3):156-160, (*2020).