Data Lakes for Digital Humanities

Jérôme Darmont
Cécile Favre
Sabine Loudcher
jerome.darmont@univ-lyon2.fr
cecile.favre@univ-lyon2.fr
sabine.loudcher@univ-lyon2.fr
Université de Lyon, Lyon 2, ERIC UR 3083
Bron, France

Camille Noûs
camille.nous@cogitamus.fr
Université de Lyon, Lyon 2, Laboratoire Cogitamus
Bron, France

ABSTRACT

Traditional data in Digital Humanities projects bear various formats (structured, semi-structured, textual) and need substantial transformations (encoding and tagging, stemming, lemmatization, etc.) to be managed and analyzed. To fully master this process, we propose the use of data lakes as a solution to data siloing and big data variety problems. We describe data lake projects we currently run in close collaboration with researchers in humanities and social sciences and discuss the lessons learned running these projects.

CCS CONCEPTS

• **Information systems** → **Data management systems**; **Database administration**; **Information integration**; • **Applied computing**; • **Digital Humanities**;

KEYWORDS

Data Lakes, Digital Humanities, Metadata

ACM Reference Format:
Jérôme Darmont, Cécile Favre, Sabine Loudcher, and Camille Noûs. 2020. Data Lakes for Digital Humanities. In *ddh20: Data and Digital Humanities 2020, October 15–17, 2020, Hammamet, Tunisia.* ACM, New York, NY, USA, 4 pages. https://doi.org/

1 INTRODUCTION

Traditional data management has long been adopted by many researchers involved in Digital Humanities (DH). However, it requires a substantial investment in data modeling, including, at the physical level, technologies such as relational and semi-structured Database Management Systems (DBMSs), various data formats, e.g., XML and JSON for semi-structured data, RDF for linked data, and query languages such as SQL and XQuery. This investment in computer science and the fact that initial data are inevitably transformed are presumably impediments to the adoption of DBMSs and related digital tools for DH.

ddh20, October 15–17, 2020, Hammamet, Tunisia

Moreover, most source information exploited by humanities and social sciences comes in textual format. Again, such textual documents are difficult to manage without substantial transformations: digitization, encoding and tagging, e.g., via the Text Encoding Initiative (TEI), and even lowercasing, stemming, lemmatization, stopword removal or normalization when it comes to text mining and natural language processing.

Another important methodological issue is the black box effect that occurs when resorting to computer scientists only "as a service". How can DH researchers work without mastering the whole process? Furthermore, designing and managing such processes also lead to research issues for computer scientists.

To leverage the above-mentioned issues, we propose the use of data lakes, a concept introduced by Dixon in 2010 as a solution to data siloing and big data variety problems [2]. Even if data exploited by DH are not always big data in terms of volume, they can bear considerable variety, i.e., including structured and semi-structured data, as well as unstructured data such as texts, various types of images, sounds and videos. Traditional data management tends to manage such heterogeneity with different systems, thus separating data into so-called silos.

A data lake is a scalable storage and analysis system for data of any type, retained in their *native format* and used *mainly* (but *not only*) by data specialists (statisticians, data scientists or analysts) for knowledge extraction [10].

One of the main advantages of data lakes is that data are stored in their initial form, and are thus recognizable by their producers, such as DH researchers. A data lake does not propose a new data model nor new data formats for data archiving. Moreover, when data are transformed for processing, the data lineage is stored as metadata, thus enforcing traceability.

However, a drawback is that unprepared data are difficult to process and require data specialists who can program. Yet, we strongly advocate, with other researchers, for the "industrialization" of data lakes, i.e., providing a software layer that allows non-data scientists such as DH researchers to transform and analyze their own data in autonomy, just as dynamic reports are prepared on top of data warehouses for the use of business (i.e, non technical) users.

The remainder of this paper is organized as follows. In Section 2, we describe data lake projects we currently run in close collaboration with researchers in social sciences and humanities. In Section 3, we conclude this paper by discussing the lessons learned running these projects.

2 EXAMPLE DH PROJECTS INVOLVING DATA LAKES

2.1 HyperThesau

The "Hyper thesaurus and data lakes: Mine the city and its archaeological archives" (HyperThesau) project involves a multidisciplinary team consisting of two research laboratories of archaeology and of computer science, a digital library, two archaeological museums and a private company. This project has two main objectives:

(1) the design and implementation of an integrated platform to host, search, share and analyze archaeological data;
(2) the design of a domain-specific thesaurus taking the whole archaeological data lifecycle into account, from data creation to publication.

Archaeological data may bear many different types, e.g., textual documents, images (photographs, drawings...), sensor data, etc. Moreover, similar documents, e.g., excavation reports, are often created by various software tools that are not compatible with each other. The description of an archaeological object also differs with respect to users, usages and time. Such a variety of archaeological data induces many scientific challenges related to storing heterogeneous data in a centralized repository, guaranteeing data quality, cleaning and transforming the data to make them interoperable, finding and accessing data efficiently and cross-analyzing the data with respect to their spatial and temporal dimensions.

To overcome all these challenges, we implement a data lake. Our approach aims to collect all types of archaeological data, save them inside the data lake and propose metadata for better organizing data and for allowing users to easily find data for analysis purposes. Our data lake prototype is architectured in nine layers (Figure 1) [5, 6].

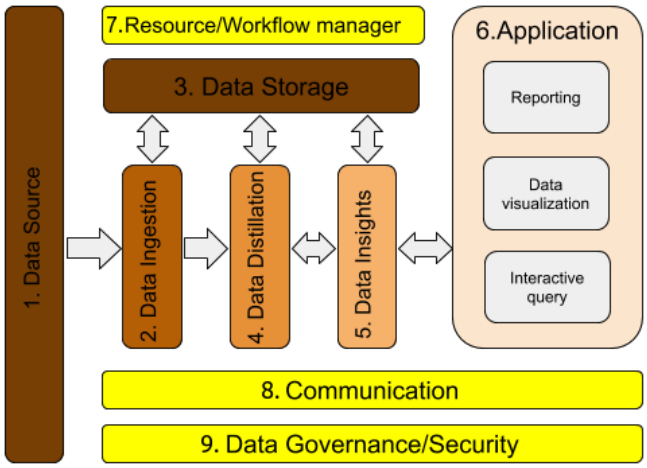

Figure 1: HyperThesau data lake's layered architecture [5]

(1) The data source layer gathers the basic properties of data sources, e.g., volume, format, velocity, connectivity, etc. Based on these properties, data engineers can determine how to import data into the lake.
(2) The data ingestion layer provides a set of tools for performing batch or real-time data integration. Data engineers can choose the right tools and plans to ingest data into the data lake with respect to data source properties and the lake's capacity. During ingestion, metadata provided by data sources, e.g., name of excavation sites or instruments, must be gathered as much as possible.
(3) The data storage layer is core to a data lake. It must have the capacity to store all data in any format.
(4) The data distillation layer provides a set of tools for data cleaning (eliminating errors such as duplicates and type violations) and encoding formalization (converting various data and character encoding).
(5) The data insights layer provides a set of tools for data transformation (e.g., into models) and exploratory data analysis (e.g., pattern discovery). Note that transformed data may also be stored into the lake for later reuse.
(6) The data application layer provides applications that allow users extracting value from data, e.g., through an interactive query system, reports or dataviz.
(7) The workflow manager layer provides tools to automate the flow of data processes.
(8) The communication layer provides tools that allow the other layers to communicate with each other. It must provide synchronous and asynchronous communication capability.
(9) The data governance layer provides a set of tools to establish and execute plans and programs for data quality control [4].

Each of the above layers is implemented with one or more frameworks of the Apache Hadoop ecosystem, e.g., Atlas[1], HDFS[2], HIVE[3], OpenLdap[4], Spark[5], etc. This prototype is operational and currently hosts the data of two archaeological research facilities. The metadata management system instantiates the MEtadata model for Data Lakes (MEDAL), which adopts a graph model [10]. It is implemented with Apache Atlas, which can host not only descriptive metadata, but also several thesauruses. With the help of a search engine, i.e., Solr[6], users can find data through descriptive metadata, a thesaurus or the data lineage.

2.2 Bretez/STRATEGE

Bretez [7] is a multidisciplinary project aiming at a visual and sonorous restitution of the XVIIIth-century Paris. It is also an exploratory project constituted of successive, interlinked modules that are (and must be) interoperable and open. The historical urban restitution is achieved with video game engines that bear their own respective characteristics, of course related to gaming. Yet, here, they are used for specific management and traceability needs. Moreover, Bretez' documentation is a voluminous corpus of heterogeneous and multimedia data.

Within project Bretez, the "Traceability and information management system for multimedia data" (STRATEGE) aims at designing, storing, querying and analyzing all the project's data. To master data heterogeneity, manage data quality and volume, warrant data interoperability and an efficient access while keeping data in their

[1] https://atlas.apache.org
[2] hhttps://hadoop.apache.org/docs/r1.2.1/hdfs_user_guide.html
[3] https://hive.apache.org
[4] https://www.openldap.org
[5] http://spark.apache.org
[6] https://lucene.apache.org/solr/

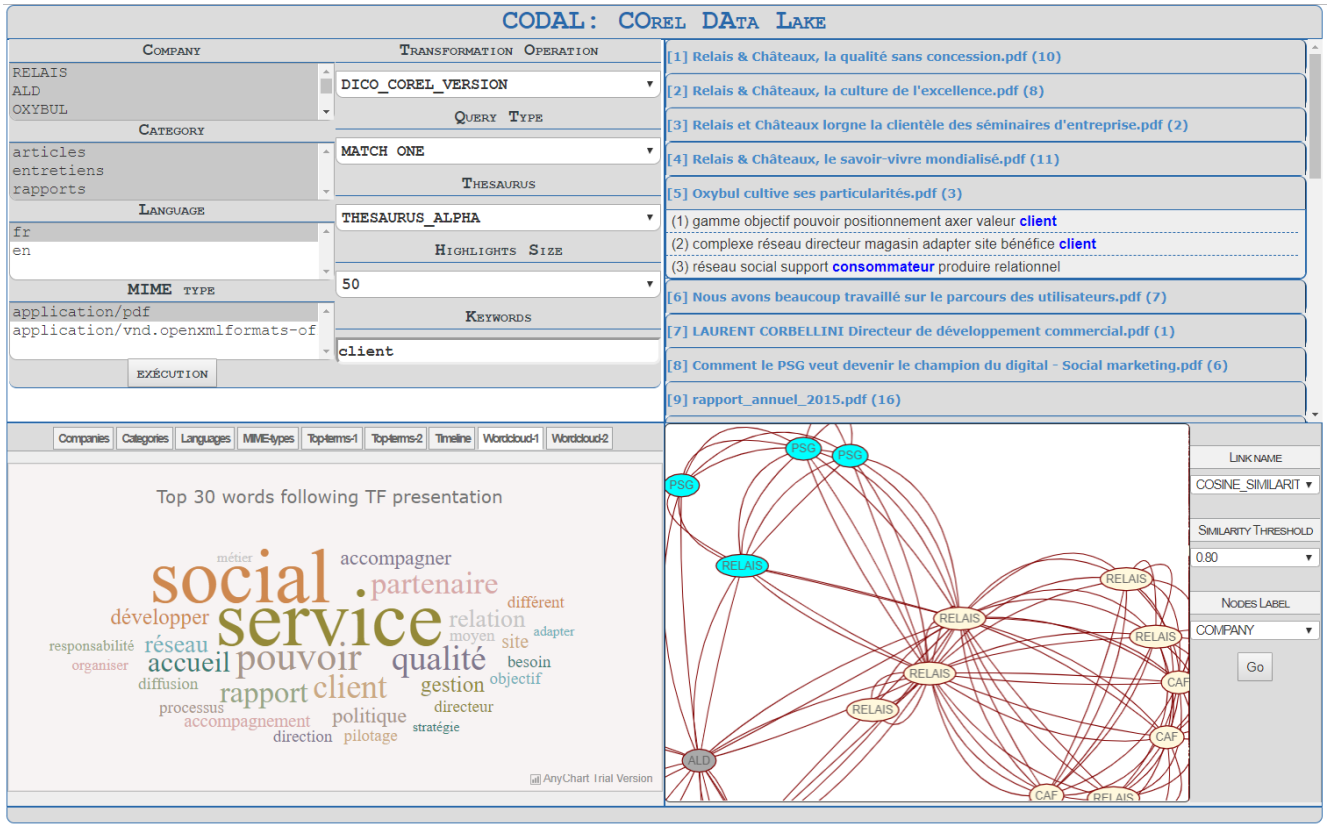

Figure 2: CODAL example screenshot [9]

original form so that they remain usable reference for the project's researchers, we resort to a data lake.

STRATEGE is in its first stages: we catalogued all existing data, which included a database, textual documents, sounds, images and a 3D Unity[7] (a game engine) model, and more. The database is particularly interesting, for it contains both data and metadata. While retaining it, we also restructured it so as to allow its metadata to interoperate with specific data lake metadata. In short, there are "business" metadata and technical metadata.

The remaining tasks include fully designing and integrating the metadata system, on the basis of MEDAL [10]; make the data from the Unity model accessible into the lake; formalize analysis needs; and design tools that must jointly handle textual, visual and audio content, as well as the heterogeneity of data sources. Such software tools must be accessible to all researchers involved in the Bretez project.

2.3 COREL and AURA PMI

Both the projects "At the heart of customer relationship" (COREL) and "Digital transformation, servicization and mutations of industrial SME business models" (AURA PMI) relate to management sciences and are carried out in collaboration with the Coactis laboratory[8]. Although their respective focus and scope are different, they are quite similar in terms of data: a corpus of various textual documents (e.g., annual reports from companies and organizations, interviews of senior or top executives, press articles; all in French or English) and data from various sources, including the Web, curated and inferred by researchers in management sciences from company legal information and performance indicators such as workforce, annual revenue, stock-exchange price and perceived level of digitization and servicization.

With such data handy, the objective is to cross-analyze the terms and expressions found in textual resources with structured, qualitative and quantitative data, in order to discover new insights regarding how companies communicate *vs.* their actual customer relationship management strategy (for COREL) and how digitization and servicization impact economic performance (for AURA PMI). The challenges here are to:

(1) leverage metadata that allow querying the whole corpus;
(2) jointly analyze structured and unstructured data;
(3) allow management science researchers to perform analyses by themselves.

To complete these tasks for the COREL project, we designed a metadata system that prefigured MEDAL [10] and proposed the lightweight COREL Data Lake architecture (CODAL) [9], which is composed of:

[7] https://unity.com
[8] http://coactis.org

(1) a storage layer that notably includes Elasticsearch[9] for indexing textual contents;
(2) a metadata layer leveraging and extending the Metadata Encoding & Transmission Standard (METS) [11], stored in the BaseX XML DBMS[10];
(3) an analysis layer, i.e., an intuitive Web-based graphical interface that allows management science researchers to perform analyses in autonomy, thus enforcing the "industrialization" of CODAL.

The analysis layer features three kinds of analyses:

(1) data exploration akin to On-Line Analytical Processing (OLAP)[1];
(2) proximity analyses such as similarity (what documents are similar or different [8]) and centrality (to identify the documents bearing a specific or common vocabulary, hinting at its importance [3]) analyses;
(3) custom highlights of the context of terms and, optionally, their synonyms, in textual documents.

All three types of analyses come with various dataviz (Figure 2).

The AURA PMI Data Lake (AUDAL) is currently being developed, and builds upon CODAL. Its metadata system will notably be a substantial evolution of MEDAL supported by the Neo4J[11] graph DBMS. Moreover, the AUDAL analysis layer, which lays on an Application Programming Interface (API), will be much more elaborate and efficient than CODAL's.

3 CONCLUSION

In all four data lake projects summarized in Section 2, we use different versions of the MEDAL metadata system, which is designed to be generic. However, although MEDAL is quite flexible, we do not believe in a single model for data lakes. There are indeed significant differences in data in only four projects, in terms of volume, variety and velocity, which imply different architectures and technologies. Thus, we think that much needed methodological tools for data lakes should be *instantiated* for each project rather than applied "as is".

Furthermore, the software layer we add to "industrialize" our data lakes might become yet another black box, while there is a strong stake for researchers in humanities and social sciences involved in DH projects not to be dispossessed of data by an analysis layer that would adopt a "click and go" approach. Data are indeed often partly constructed by said researchers themselves as a product of scientific work that takes time, thus giving a significant value to datasets.

In consequence, we take great care of accompanying DH users in their appropriation of our analysis tools, not only by training, but especially by interweaving research methodologies from computer science and other disciplines *by design*, in close collaboration with partner researchers.

Moreover, the possibility of having both access to the raw data and the entire possible processing chain is necessary, because black boxes are seldom compatible with a sound methodological approach aiming at producing scientific knowledge. Data lakes precisely allow this much needed transparency.

ACKNOWLEDGEMENTS

Projects HyperThesau and Bretez/STRATEGE are funded by the Laboratory of Excellence "Intelligence of Urban Worlds" (IMU)[12]. Project COREL was funded by the University of Lyon 2. Project AURA PMI is funded by the Auvergne-Rhône-Alpes Region.

REFERENCES

[1] E.F Codd, S.B Codd, and C.T Salley. 1993. *Providing OLAP (On-line Analytical Processing) to User-Analysts: An IT Mandate.* Technical Report. E.F. Codd & Associates.
[2] James Dixon. 2010. Pentaho, Hadoop, and Data Lakes. (October 2010). https://jamesdixon.wordpress.com/2010/10/14/pentaho-hadoop-and-data-lakes/
[3] Ashley Farrugia, Rob Claxton, and Simon Thompson. 2016. Towards Social Network Analytics for Understanding and Managing Enterprise Data Lakes. In *Advances in Social Networks Analysis and Mining (ASONAM 2016), San Francisco, CA, USA (IEEE)*. 1213–1220.
[4] Vijay Khatri and Carol Brown. 2010. Designing data governance. *Commununications of the ACM* 53 (January 2010), 148–152.
[5] Pengfei Liu, Sabine Loudcher, Jérôme Darmont, and Camille Noûs. 2020. *A data lake architecture and implementation for archaeological data management and analytics.* Technical Report. Université de Lyon, Lyon 2, ERIC UR 3083, France.
[6] Pengfei Liu, Sabine Loudcher, Jérôme Darmont, Emmanuelle Perrin, Jean-Pierre Girard, and Marie-Odile Rousset. 2020. Metadata model for an archeological data lake. Digital Humanities (DH 2020), Ottawa, Canada. (July 2020). https://dh2020.adho.org
[7] Mylène Pardoën. Last accessed: June 2020. Bretez site officiel. (Last accessed: June 2020). https://sites.google.com/site/louisbretez/home
[8] Pascal Pons and Matthieu Latapy. 2006. Computing Communities in Large Networks Using Random Walks. *Journal of Graph Algorithms and Applications* 10, 2 (2006), 191–218.
[9] Pegdwendé-Nicolas Sawadogo, Tokio Kibata, and Jérôme Darmont. 2019. Metadata Management for Textual Documents in Data Lakes. In *21st International Conference on Enterprise Information Systems (ICEIS 2019), Heraklion, Crete-Greece.* INSTICC, 72–83. Vol. 1.
[10] Pegdwendé-Nicolas Sawadogo, Etienne Scholly, Cecile Favre, Eric Ferey, Sabine Loudcher, and Jérôme Darmont. 2019. Metadata Systems for Data Lakes: Models and Features. In *1st International Workshop on BI and Big Data Applications (BBIGAP@ADBIS 2019), Bled, Slovenia (Communications in Computer and Information Science)*, Vol. 1064. Springer, Heidelberg, Germany, 440–451.
[11] The Library of Congress. 2017. METS: An Overview and Tutorial. (March 2017). http://www.loc.gov/standards/mets/METSOverview.v2.html

[9] https://www.elastic.co/elasticsearch/
[10] http://basex.org
[11] https://neo4j.com
[12] https://imu.universite-lyon.fr