Title: Metadata model for an archeological data lake

Authors: Pengfei Liu, Sabine Loudcher, Jérôme Darmont, Emmanuelle Perrin, Jean-Pierre Girard, Marie-Odile Rousset

Keywords: Archaeology, Thesaurus, Data lake, Metadata

The HyperThesau project was initiated by a multidisciplinary team consisting of two research laboratories of archaeology and computer science, respectively, a digital library, two archeological 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.
Archeological data may bear many different types, e.g., textual documents, images (photos, 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 (De Luca et al., 2016). Such variety of archeological 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 by taking their spatial and temporal dimensions into account.

To overcome all these challenges, we exploit the concept of data lake (Dixon, 2010). A data lake is a centralized repository that helps store all types of data without having any predefined structure (Hai et al., 2016). To avoid a data lake turning into an inoperable data swamp, an efficient metadata management system is essential to catalogue, search and access the data (Alrehamy & Walker, 2015; Hai et al., 2016). Note that we do not propose a new archaeological data model nor new data formats for data archiving here. Our approach aims to collect all types of archaeological data, save them inside a data lake and propose metadata for better organizing data and for allowing users to easily find data for analysis purposes.

In this article, we first present a data lake prototype that is architectured in nine layers such as data ingestion, data storage, data application, data governance, data security, etc. Each layer is implemented with one or more frameworks of the Hadoop ecosystem, e.g., Atlas, HDFS, HIVE, OpenLdap, Spark, etc. This prototype is operational and currently hosts the data of two archeological research facilities. Secondly, we present our metadata management system and the metadata model we use to manage archeological data inside the data lake. We actually instantiate the MEtadata model for Data Lakes (MEDAL), which adopts a graph model (Sawadogo et al., 2019). Eventually, our metadata management system 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, users can find data through descriptive metadata, a thesaurus or the data lineage.

References

Alrehamy H. et Walker C. *Personal Data Lake With Data Gravity Pull*. 5th International Conference on Big Data and Cloud Computing(BDCloud 2015), Dalian, China, Volume 88 of IEEE Computer Society Washington, pp. 160–167. 2015.

De Luca L. et al., *Patrimonialisation du numérique et numérisation du patrimoine, regards croisés*. MICNRS, 2016. Consultable à l'adresse : www.cnrs.fr/mi/IMG/pdf/synthesepatrimoine_numeriquevf.pdf

Dixon J. *Pentaho, Hadoop, and Data Lakes*. 2010. Consultable à l'adresse : https ://jamesdixon.wordpress.com/2010/10/14/pentaho-hadoop-and-data-lakes/

Hai R., Geisler S. et Quix C. *Constance : An Intelligent Data Lake System*. International Conference on Management of Data (SIGMOD). San Francisco, USA, ACM Digital Library, pp. 2097–2100. 2016.

Sawadogo P., Scholly E., Favre C., Ferey E., Loudcher S., Darmont J. *Metadata Systems for Data Lakes: Models and Features*. 1st International Workshop on BI and Big Data Applications (BBIGAP@ADBIS 2019), Sept 2019, Bled, Slovenia. pp.440-451. 2019.