



Archaeology in the Digital Age: From Paper to Databases

Frédérique Mélanie--becquet, Johan Ferguth, Katherine Gruel, Thierry Poibeau

► To cite this version:

Frédérique Mélanie--becquet, Johan Ferguth, Katherine Gruel, Thierry Poibeau. Archaeology in the Digital Age: From Paper to Databases. Digital Humanities 2015, Jun 2015, Sydney, Australia. 2015, Proceedings of the conference "Digital Humanities 2015". <hal-01173964>

HAL Id: hal-01173964

<https://hal.archives-ouvertes.fr/hal-01173964>

Submitted on 8 Jul 2015

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Archaeology in the Digital Age: From Paper to Databases

Frédérique Mélanie-Becquet*, Johan Ferguth*, Katherine Gruel** & Thierry Poibeau*

*Laboratoire LATTICE

CNRS & Ecole normale supérieure & Université Paris 3 Sorbonne Nouvelle

**Laboratoire AOROC

CNRS & Ecole normale supérieure & Ecole Pratique des Hautes Etudes

45 rue d'Ulm 75005 Paris

firstname.lastname@ens.fr

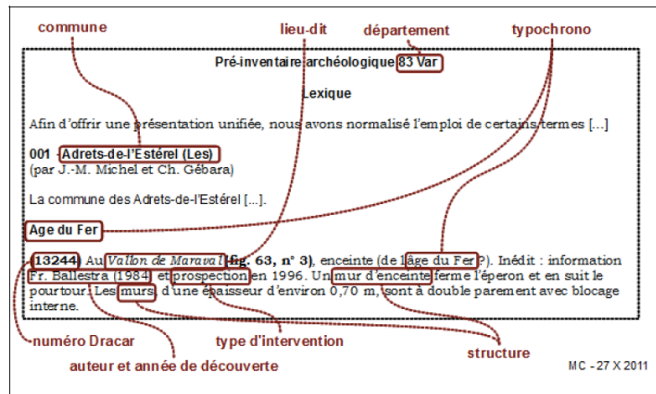
Research units in archaeology often manage large and precious archives containing various documents, including reports on fieldwork, scholarly studies and reference books. These archives are of course invaluable, recording decades of work, but are generally hard to consult and access. In this context, digitizing full text documents is not enough: information must be formalized, structured and easy to access thanks to friendly user interfaces.

The situation at AOROC, a research unit of Ecole normale supérieure specialized in archaeology, is precisely the one described hereabove: several decades of research are contained in documents, which are hardly accessible, even for people working in the lab. The situation is such that researchers may produce studies largely overlapping with previous work, which remained unknown because of its poor accessibility.

A partnership has thus been established between AOROC and LATTICE, another research unit specialized in Digital humanities and natural language processing, to digitize and structure a part of these documents. A pilot study concerned a collection of texts covering excavations related to the Gaul period (*Cartes Archéologiques de la Gaule*, [Provost, 1988-]), over an area encompassing a large part of modern France from the Iron Age to the Medieval period (800BC to 800AD). 128 volumes have been published so far: each volume corresponds to one French department (some departments are covered by several volumes). The pilot study concerned three of these volumes, along with other types of documents so as to ensure the genericity of the developed solution. The idea is of course not just to digitize and transfer documents online but also to extract key information so as to feed structured databases (Poibeau et al., 2013). The result should be accessible using a standard but powerful query language.

A first step consists in recognizing the structure of the documents, which mainly consist in notices, each notice corresponding to a specific “municipality” (the structure is not formally encoded in the source documents and may vary from one document to the other). Specific zones inside the notices have to be recognized (see **figure 1**): this can be done by specific scripts but also needs some manual cleaning. In our opinion, the most interesting part concerns the natural language processing techniques used for information extraction. These include:

- Named entity recognition (i.e. the recognition of proper names, location, dates, etc.)
- Technical term extraction (i.e. all archaeological terms)
- Entity linking (i.e. the recognition of the different variants of a same term or entity and its connection to the same type of object in the database).



Different tools have been used, like TreeTagger (<http://www.cis.uni-muenchen.de/~schmid/tools/TreeTagger/>) for part-of-speech tagging, Yatea (<https://metacpan.org/release/Lingua-YaTeA>) for technical term extraction and TyDI for terminology structuration (<http://ciam.inra.fr/logiciels/node/195>). In-house solutions have been developed for document structure, named entity recognition and entity linking. All the modules are parameterized so that they can be easily adapted to new sources of data. These tools are based on up-to-date natural language processing techniques and have obtained excellent results in recent benchmarks like Semeval (Ruiz and Poibeau, 2015).

The result of the project made it possible to automatically feed a structured database based on the textual content analysis. Documents are now accessible online with full text facilities, structured indexes and ontologies (see **figure 2**). It is thus possible to interrogate the database with queries dealing with a specific location, a specific series of objects or a given period of time.

	008 - Saint-Céols (I.N.S.E.E. n° 202)	structure → construction → Evacuation des eaux → aqueduc /moblier → art /structure → dépôt → bâtiment /moblier → récipient → céramique /département → Cher /moblier → objet domestique → coureau /structure → dépôt /moblier → récipient → dépotoir → dépôt monétaire /structure → funéraire → enclos /structure → dépôt → enclos quadrangulaire /moblier → récipient → céramique → vase → pot /moblier → récipient → seau /Ecofat → silex /structure → construction → sol /moblier → trou /structure → habitat → village gallo-romains → villa romaine /structure → village /structure → voie romaine → voie	008 - Saint-Céols (I.N.S.E.E. n° 202) Aucun site archéologique n'est mentionné dans la bibliographie.
GERARD COLON <pre> %<racine %<departement preprofondeur="0" nom="departement" pere="father">...</dep %<epoque protohistorique preprofondeur="1" nom="epoque protohistorique %<age du bronze preprofondeur="2" nom="Age du Bronze" pere="epoque pr %<age du bronze final preprofondeur="3" nom="Age du Bronze final" p %<fin de l'âge du bronze preprofondeur="3" nom="fin de l'Age du %<age du bronze %<age du fer preprofondeur="2" nom="Age du Fer" pere="epoque protohisti %<debut du premier âge du fer preprofondeur="3" nom="debut du premi %<fin du premier âge du fer preprofondeur="3" nom="fin du premier âge du %<fin de l'âge du fer preprofondeur="3" nom="fin de l'Age du %<omission du premier siècle av preprofondeur="3" nom="omission du %<omission du premier siècle av preprofondeur="3" nom="debut du premi %<age du fer> %<epoque du hallstatt preprofondeur="2" nom="epoque du hallstatt" per %<debut du hallstatt ancien preprofondeur="3" nom="debut du hallsta %<hallstatt ancien preprofondeur="3" nom="hallstatt ancien" pere="e %<epoque du hallstatt> %<la tene preprofondeur="2" nom="La Tène" pere="epoque protohistoriq %<tene ancienne preprofondeur="3" nom="Tène ancienne" pere="La Tène %<tene moyenne preprofondeur="3" nom="Tène moyenne" pere="La Tène"> %<tene finale preprofondeur="3" nom="Tène finale" pere="La Tène"> %<epoque de la guerre des gaules preprofondeur="4" nom="epoque de %<tene finale> %</la tene> %<epoque protohistorique> %<epoque medieval preprofondeur="1" nom="epoque medieval" pere="chronologie"> %<epoque age preprofondeur="2" nom="Moyen Age" pere="epoque medieval"> %<debut du moyen age preprofondeur="3" nom="debut du Moyen Age">=<debut du moyen age> %<haut moyen age preprofondeur="3" nom="haut Moyen Age" pere="Moyen Age">=<haut_moyen_age> %<moyen age> %<epoque merovingienne preprofondeur="2" nom="epoque merovingienne" pere="epoque medieval">=<epoque merovingienne> %<epoque carolingienne preprofondeur="2" nom="epoque carolingienne" pere="epoque medieval">=<epoque carolingienne> %<epoque medieval> %<epoque chassienne preprofondeur="1" nom="epoque chassienne" pere="chronologie">=<epoque chassienne> %<epoque moderne preprofondeur="1" nom="epoque moderne" pere="chronologie">=<epoque moderne> %<epoque gallo-romaine preprofondeur="1" nom="epoque gallo-romaine" pere="chronologie"> %<debut de l'epoque romaine preprofondeur="2" nom="debut de l'epoque romaine" pere="epoque gallo-romaine"> %<epoque librecleuse preprofondeur="3" nom="epoque Librecleuse" pere="debut de l'epoque romaine"> %<moitie du premier siecle preprofondeur="4" nom="moitie du premier siecle" pere="epoque librecleuse">=<moitie du premier siecle> %<epoques augustoclaudiennes preprofondeur="3" nom="epoques augustoclaudiennes" pere="debut de l'epoque romaine"> %<epoque augustienne preprofondeur="4" nom="epoque augustienne" pere="epoques augustoclaudiennes"> %<debut de l'epoque augustienne preprofondeur="5" nom="debut de l'epoque augustienne" pere="epoque augustienne">=<debut de l'epoque augustienne> %<quart du premier siecle preprofondeur="5" nom="quart du premier siecle" pere="epoque augustienne">=<quart du premier siecle> %<epoque augustienne> %<epoques augustoclaudiennes> %<epoque claudienne preprofondeur="3" nom="epoque claudienne" pere="debut de l'epoque romaine">=<epoque claudienne> %<epoque tibérienne preprofondeur="3" nom="epoque tibérienne" pere="debut de l'epoque romaine">=<epoque tibérienne> %<debut de l'epoque romaine> </pre>			008 - Saint-Germain-du-Puy (I.N.S.E.E. n° 213) Cette commune est traversée d'est en est par la voie Bourges-Sancerre : A. Buhot de Kersers, Statistique monumentale, 1875, p. 48 ; M.-E. Fonvieille, J. Deguérat, dans Cahiers du Berry, 60, 1920, p. 21-26. Aux Boudards (fig. 15), sur la terrasse alluvionnaire, dans la vallée du Langis, J. Holmgren a découvert, en 1976, "une vaste villa gallo-romaine, avec cour centrale (d'au moins 180 m de long) et bâtiment principal à l'est ; le matériel céramique (sigillée, commune et Argonne) date des IIe au IVe s. : A. Ledy, La campagne, 1800 (I), p. 53, 64, 69, 80, 88, 101, 108, 110 (II), (II), p. XII ; J. Holmgren, A. Ledy, dans Cahiers du Berry, 60, 1980, p. 9, II-3 ; V. de Kisch, dans Gallia, 36, 2, 1980, p. 317 (prospection aérienne et au sol J. Holmgren et A. Ledy) ; - O. Ruffier, dans Cahiers du Berry, 103, 1990, p. 63-64 ; J. Holmgren, site 122. A Tierceville, J. Holmgren a photographié un enclos quadrangulaire : J. Holmgren, site 330. Au lieu-dit Le Champ des Pierres ou le Domaine Neuf, près de Turly, on a trouvé, en 1867, des débris de tuiles à rebords, et quelques monnaies romaines du Haut-Empire (Domitian, Hadrien, Antonin, Marc-Aurèle, Gallien) : A. Buhot de Kersers, Statistique monumentale, 1875, p. 47 ; - L. Martinet, Le Berry préhistorique, 1878, p. 101 ; - O. Ruffier, dans Cahiers du Berry, 103, 1990, p. 63. Au Domaine des Augustins, un antoninien de Claude II :

This work can be compared to other initiatives with a similar goal. In archaeology, one can cite archaeological archiving bodies such as the ADS (<http://archaeologydataservice.ac.uk/>), tDAR

(<http://www.tdar.org/about/>), or OpenContext (<http://opencontext.org/>), among others. Research on interoperability between archaeological databases includes, among many others (Binding et al., 2008; Doerr, 2003; Jordal et al., 2004; Vlachidis and Tudhope, 2012). Our project is different since it is from the beginning designed to deal with texts in different languages, especially French, German and English, with a cross-linguistic perspective. One of the major research issues is the maintenance of an international terminology referring to complex notions that can vary from one country to the other. The system should be flexible enough to be able to match related concepts (even if they vary slightly from one source to the other), but relevant enough so as to provide only relevant information. This goal involves a permanent dialogue between experts of the domain and the maintenance of an up-to-date terminology. The tools and interfaces developed within the project should help to keep this goal a reality as much as possible.

Acknowledgements

This work has received support of Paris Sciences et Lettres (program “Investissements d’avenir” ANR-10-IDEX-0001-02 PSL*) and of the laboratoire d’excellence TransferS (ANR-10-LABX-0099). The project has been realised during the PEPS CNRS-PSL EITAB 2013-2014.

References

- Binding, C., May, K. and Tudhope D. (2008). Semantic Interoperability in Archaeological Datasets: Data Mapping and Extraction Via the CIDOC CRM. In Research and Advanced Technology for Digital Libraries. Berlin: Springer (LNCS 5173), pp 280-290.
- Doerr, M. (2003). The CIDOC Conceptual Reference Module: An Ontological Approach to Semantic Interoperability of Metadata. AI magazine, Vol. 24 n°3, pp 75-92.
- Jordal, E., Holmen J. and Olsen S. A. (2004). From XML-tagged Acquisition Catalogues to an Event-based Relational Database. In F. Niccolucci (ed.), Beyond the artefact - Proceedings of CAA2004, Archaeolingua, pp 81-85.
- Provost M. (ed., 1988-). Cartes archéologiques de la Gaule. Paris: Académie des Inscriptions et Belles Lettres / FMSH editions, 128 volumes published to date.
- Poibeau, T., Saggion, H., Piskorski J. and Yangarber R. (eds. 2013). Multi-source, Multilingual Information Extraction and Summarization. Berlin: Springer (Theory and Applications of Natural Language Processing).
- Ruiz P. and Poibeau T. (2015). Entity Linking Combining Open Source Annotators via Weighted Voting. Boulder : Proceedings of Semeval 2015.
- Vlachidis A. and Tudhope D. (2012). A pilot investigation of information extraction in the semantic annotation of archaeological reports. International Journal of Metadata, Semantics and Ontologies archive. Vol. 7 Issue 3, pp. 222-235.