Olivier Marlet, Xavier Rodier

A way to express the reliability of archaeological data: Data traceability at the *Laboratoire Archéologie et Territoires* (Tours, France).

Affiliation: UMR 7324 CITERES-LAT, CNRS, Université de Tours.

Address: UMR 7324 CITERES-LAT, BP 60449, 37204 TOURS cedex 03

Corresponding author: olivier.marlet@univ-tours.fr / +332 47 36 15 06

Olivier Marlet ORCID : 0000-0001-9422-1418

Xavier Rodier ORCID : 0000-0002-1243-3167

Abstract

In order to respect the good practices in archaeology disseminated by the MASA Consortium (Archaeologists and Archaeological Sites Memories), the *Laboratoire Archéologie et Territoires* (Tours, France) wished to evaluate the progress of ArSol database (Soil Archives), its field data management database, with regard to the FAIR principles or the 5 Stars Linked Open Data. The work undertaken to achieve compliance with these precepts has shown that it is also necessary to ensure the relevance and reliability of the published data.

For data to be reusable, it seems essential to ensure traceability. Various tools set up for the ArSol database make it possible to ensure this traceability from the field recording, through its exploitation, to the publication of the results of the excavation.

The traceability of data, to ensure their quality in terms of reliability and relevance, is an aspect that is fortunately already taken into account in the ArSol database and that complements satisfactorily the FAIR principles requirements.

Keywords: archaeology, database, traceability, FAIR principles, LOD, interoperability

International Journal on Digital Libraries (Springer-Verlag GmbH Germany)

https://doi.org/10.1007/s00799-019-00274-3

Received: 28 January 2019 / Revised: 15 July 2019 / Accepted: 29 July 2019

Introduction

Thanks to the involvement of the *Laboratoire Archéologie et Territoires* (LAT) in the Consortium *Mémoires des Archéologues et des Sites Archéologiques* (MASA), the ArSol online archaeological database (*Archives du Sol*) is improving to meet the requirements of the linked open data. In recent years, important work has already been done on metadata, alignment of vocabularies with a standardized thesaurus (PACTOLS) and data interoperability by mapping with the ontology of the CIDOC CRM. The remaining tasks are assessed through good practices that have been spreading more and more within the Digital Humanities communities in the last few years, and more recently in archaeology. These good practices are defined by the 5 stars Linked Open Data [3] and the FAIR principles [5]. Thus, we know quite precisely what tasks we still have to do to ensure that our data comply with these good practices. However, all these efforts are effective only if the data we wish to disseminate are relevant and reliable. How can we guarantee users the relevance and reliability of the data we make available?

A first way to ensure the reliability and relevance of the data is to guarantee that, from the field to the database, the information is not degraded during the computerization process. A second way is to ensure the methodological transparency of statistical treatments, such as those used to date an archaeological context using the pottery found. Finally, a third way is to formalize scientific publications in archaeology by using the logicist format to explain the arguments of a reasoning and by allowing the reader to trace back to the evidence that field data constitute [8, 9].

1 ArSol: an information system for field recording

Archaeology destroys most of its study object during excavations. Thus, the field recording becomes the primary documentation on which the archaeologist can rely after the excavation. As such, the excavation archives of Tours (France) and the parish centre of Rigny (France) are emblematic collections of urban and mediaeval archaeology. They present structured documentation based on a homogeneous stratigraphic recording system imported from the United Kingdom in Tours in the late sixties and then disseminated in France [6, 7].

The ArSol information system has been developed by the LAT to process the data of the excavations carried out by its members. First deployed in the early nineties, ArSol integrates born digital more recent data and makes it possible to link old digitised documentation. This system is used for stratigraphic excavations and has the dual purpose of data management and research. In particular, it makes it possible to manage the chronological sequences of the stratigraphic elements and to provide a phasing approach to the site [11].

Long before computerization, this recording system from the Post-classical Archaeology was very robust [8]. The stratigraphic units recorded on the field are combined into spatio-functional elements (features, including walls and burials). Stratigraphic units and features are also combined into time elements, according to the relative chronology of stratigraphic relationships. The temporal groupings are combined to form sequences or aggregations. A spatio-temporal grouping of sequences and aggregations creates sets, which are now enhanced by the datings. The interpretation of the site into periods and phases is based on the assumptions that these chronological divisions are included in a more global historical model. Archaeological interpretation meets these two approaches, both empirical-inductive and hypothetical-deductive. (Fig.1).

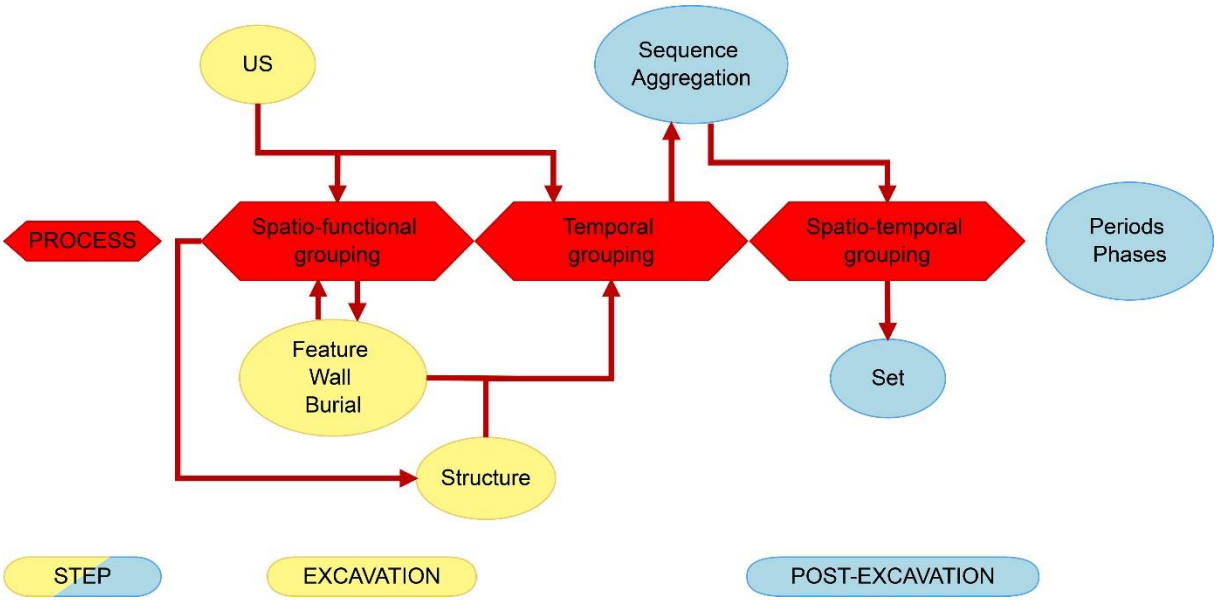


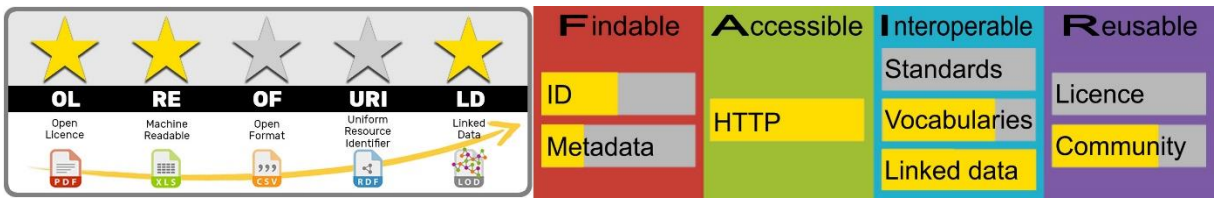
Fig.1 Diagram showing the structuring of Arsol system

Beyond being a computer system for recording and processing field data, ArSol database is the backbone of a chain of processing archaeological information from excavation to publication by ensuring the administration of evidence. ArSol database is online since 2014 (http://arsol.univ-tours.fr).

2 Assess our datasets using Five Stars LOD and FAIR principles

To improve the reliability of the data, it is not enough to simply give information about the entry of the form (date and author). The traceability of information would be more efficient by clarifying the methods for field recording, the methods for interpreting a feature or for phasing. Thus, researchers know the conditions for acquiring data or reasoning implemented. From a software point of view, it is essential to update the structure or fields of the database by updating the documentation (in particular the data dictionary), especially if the structure evolves.

In the framework of MASA consortium, we assessed the status of our datasets at our laboratory, according to the Five Stars LOD [3] and the FAIR principles [5]. This assessment shows the obvious: the older the datasets are, the less they comply with FAIR principles. The FAIR evaluation helps us to identify gaps to be filled in some of our datasets, in particular ArSol (Fig.2). For each of our datasets, here the ArSol database, we have tried to show its state of progress in the stars of the LOD and in the FAIR principles. The ArSol database is special because it is an old database and it is in a proprietary format. Despite this, it is accessible online (http://arsol.univ-tours.fr) and we were able to create an RDF export mapped with the CIDOC-CRM ontology. This is why ArSol won the Linked Data star but not the Open Format star. For the FAIR principles, a yellow progress bar indicates progress in each area.



We have therefore been able to evaluate the quality of our data according to the criteria of these two evaluation systems: sustainable identification of resources, metadata, accessibility, standard format, standardized vocabulary, data linked to gazetteers, licenses, compliance with community standards. Although Aalto University (Finland) has been working to extend the five stars to seven [13] and even if the two evaluation methods are similar, the FAIR principles method seemed more detailed and complete than the Five Stars Linked Open Data method, so we focused mainly on the FAIR principles. Using a very subjective evaluation grid, we were able to evaluate all the remaining work to improve the different datasets we publish on the web, and in particular the ArSol database.

2.1 Metadata

Within the online publication of ArSol, the paper recording sheets have been digitized so that they can be put online, combined with their computerized version. This entire digitized archive group has been accurately described in an EAD (Encoded Archival Description) file, an XML-based encoding standard for archival finding aids.

In addition to the physical description and description of the archival units constituting the collection, this metadata file specifies how the collection is organized, the classification method and elements on the recording methodology. By defining the conditions for data acquisition and the way they are structured, the user takes a critical look at the quality and reliability of the data.

2.2 Vocabulary management

ArSol does not manage thesauri but manages open vocabulary lists matched with PACTOLS standard multilingual thesauri to ensure data interoperability. The PACTOLS are initially intended for documentary indexing, but work is in progress to make them thesauri of the archaeological discipline. When the operator enters data in ArSol, he first consults the terms available in the PACTOLS thesaurus. Then, if he does not find what he looks for, he looks among the ArSol vocabulary not matched with the PACTOLS. Finally, if he still can't find it, he can add a new term to the list. Terms unmatched with PACTOLS will try as much as possible to integrate PACTOLS to improve the thesauri.

To facilitate this alignment with the PACTOLS thesauri, a web interface has been developed by Marion Lamé in the context of a Post PhD at the LAT in collaboration with Federico Ponchio (CNR, Pisa), within the framework of the MASA Consortium. The OpenTermAlign interface helps you to align a list of controlled vocabularies from a dataset with a standard thesaurus expressed in Skos. At the end of the processing, each aligned term is associated with the sustainable URL (ARK) in the thesaurus. The non-

aligned terms can then be proposed to the PACTOLS managers to enhance the thesaurus (Fig.3).

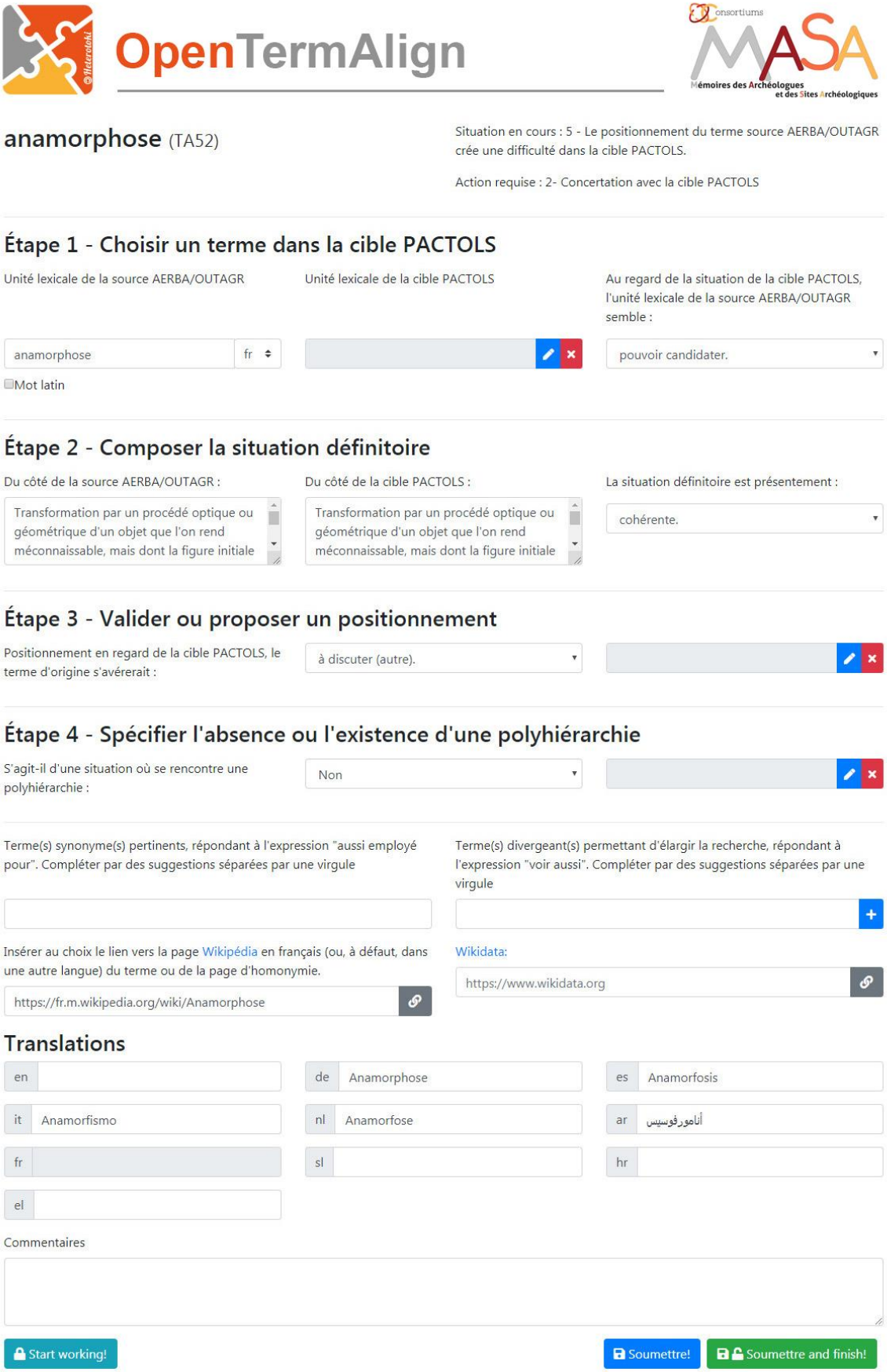


Fig.3 Prototype of OpenTermAlign, application managing the alignment of an unstructured vocabulary with a standardized thesaurus.

2.3 Database interoperability

For almost 30 years, the LAT has been using the proprietary 4D© database on which heavy developments have been made. To convert this database into a free and open format is currently unachievable due to a lack of human and financial resources. Despite the use of 4D©, non-standard format and proprietary software, we have succeeded in making ArSol data interoperable. Mapping was possible with CIDOC CRM ontology using the Protégé software (https://protege.stanford.edu/) and the Ontop (https://ontop.inf.unibz.it/) extension to connect the database with the CIDOC ontology [14]. We have set up a generic mapping of stratigraphic recording to make ArSol's main data interoperable. This mapping is currently being used in the OpenArchaeo platefom for a query interface on a SPARQL Endpoint in development by the MASA consortium.

All the work performed by the MASA Consortium on the interoperability of archaeological data is materialized through the publication of the OpenArchaeo platform [16]. The purpose of this platform is threefold. On the one hand, the archaeological datasets produced by the MASA consortium will be made available on the semantic web, in the form of a MASA triplestore aligned with the ontology of CIDOC-CRM and its extensions dedicated to archaeology, thanks to a generic model. Secondly, it aims to provide an intuitive query interface for archaeological data inspired by the ResearchSpace search engine set up by the British Museum, while OpenArchaeo is specifically dedicated to archaeology. Third, the objective is also to be able to query external triplestores at the same time as the MASA triplestore, thanks to a federation of queries. Thus, this interface queries both the MASA triplestore and any other triplestore whose data has also been mapped to the ontology of CIDOC-CRM. To optimize the matching of datasets, a generic mapping model has been implemented to link the concepts shared by most archaeological datasets, thus making them interoperable (Fig .4). This generic model can be used for any archaeological dataset. ArSol is one of the first datasets that will be accessible by OpenArchaeo.

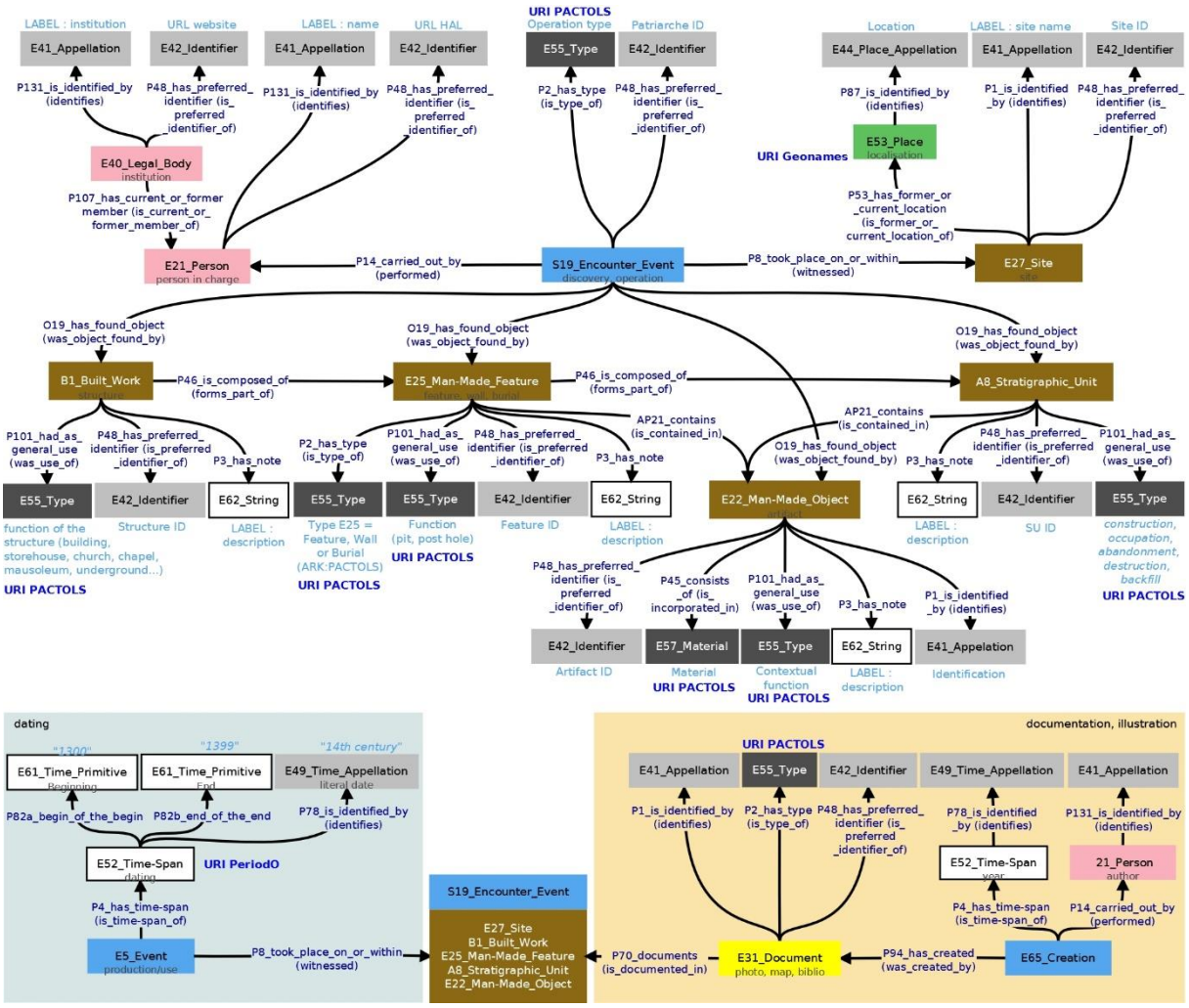


Fig.4 Archaeologic generic mapping model for OpenArchaeo with CRMbase 6.2.3, CRMsci 1.2.5, CRMarchaeo 1.4.8, CRMba 1.4

It is an important first step to make data accessible and reusable by aligning with standardized vocabularies and mapping them to an ontology recognized by the community. However, what is the benefit of publishing this data if we are not able to guarantee to those who will reuse it that this data is relevant and reliable?

3 Relevance and reliability of the ArSol data

3.1 From paper sheets to database: a loss of information?

Originally only on paper sheets (1968) [6], this record has been computerized since 1990 [7 & 11]. Already, before computerization, the sheets have become more and more formalized. However, despite this computerization, paper records are still being accumulated. First of all, because a database, with its tables and fields, its closed lists, can sometimes be less flexible than a paper sheet that can be adapted according to unforeseen cases [10], it is also very easy to delete, modify and complete the recording with a simple pencil stroke. Complete paper recording with digital recording is also a good solution to ensure a backup on different media.

From the beginning, ArSol was designed as an interoperable modular system. Thus, in addition to the field recording, one module manages specifically the pottery data and another one manages the documentation and archives related to the excavation. The input interface in the system makes is possible to specify with which module and on which datasets we want to work. It was developed with 4D©.

Computerization with closed lists and checkboxes can be considered as a coercion. Should we consider that the computerization in a database of data from paper sheet is a loss of information? The initial objective was never to duplicate the recording but to capture only information useful for computer data processing. This database constraint is also what gives it its strength since a corpus can be homogenized and the information is comparable and quantifiable, which is crucial for archaeologists (Fig.5). Can the flexibility of the paper sheet be conciliated with a structured database?

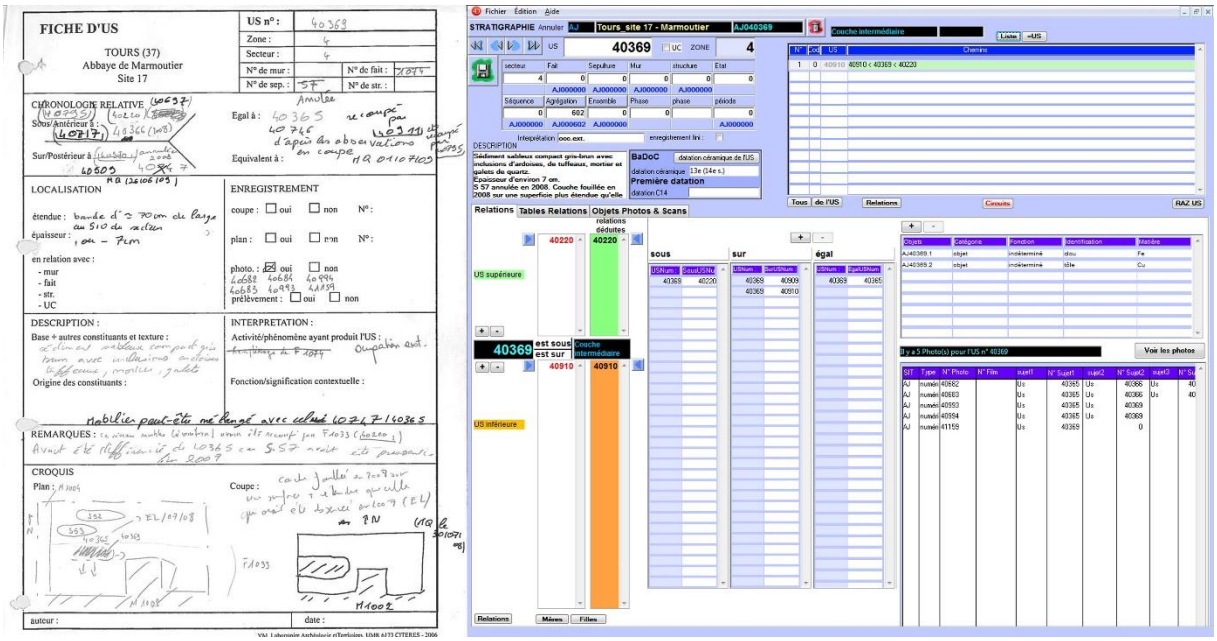


Fig.5 From paper sheet to computer for archaeological excavation of Marmoutier (France)

Thanks to the easiness of use of tablets for field recording, it becomes possible to consider a dematerialization of the recording. Then, the input interface for these tablets must be as close as possible to the field's requirements. We tested the use of tablets to make a direct digital input. The ArSol interface was then adapted to get as close as possible to the input as it is done on paper:

- The field interface provides versioning management (description, interpretation, commentary) to keep a log of the data and to follow its evolution in case of change;
- It is possible to make sketches or annotated photos;
- A free comment field is used to save field notes, data that do not necessarily fit into the other formatted fields of the database.

In ArSol, a project is underway to digitize all these recording sheets and make them accessible. They are either associated with the digital recording if it is available, or constitute the main data with only its identifier as an index.

3.2 Methodological reliability: case of pottery dating

What confidence can be placed in a date given in a database? The ArSol system has a specific module for the treatment of pottery. The examination of pottery is based on a typological reference system shared by an important network of specialists of pottery in France: ICERAMM (http://iceramm.univ-tours.fr). To provide reliable datation, the first constraint is obviously to take into account only dated artefacts found in context, not redeposited, ensuring a coherent context. Each pottery technical group (type of clay and firing) is dated according to a statistic model and a reference system ICERAMM that is refined as the data are enhanced [1, 2, 12]. In the ArSol database, all the results of this model are used to date archaeological contexts. This is the synthesis of the time range common to all the technical groups found in the context. Thus, the module does not produce a date for a set but it proposes a date range based on statistical model and pottery quantification, recalculated on the fly, taking into account the scale of the set (stratigraphic units, feature, structure). The result can therefore change if the precision of this reference frame increases.

In the interface of the online database of ArSol, an help to date a context (Fig.6) can be obtained dynamically. The result specifies the statistical process used to propose a range of dates. These details therefore allow a researcher to criticize the proposed outcome, if the researcher is able to understand these methods.

FOURCHETTE CHRONOLOGIQUE DURANT LAQUELLE TOUS LES GT PEUVENT EXISTER
(conserve le groupe le + récent si son effectif est statistiquement retenu)

1300	1379

Datation du contexte :

AA000028

EFFECTIF TOTAL DE LA SELECTION (EN NMI) 86

LE (OU LES) GT DONT L'EFFECTIF EST LE PLUS IMPORTANT (en NMI)

GT	data inf	data sup	effectif
to1j	1260	1382	39

Fourchette chronologique de l'Ensemble : estimation modèle statistique

Borne inférieure 0

Borne supérieure 0

LE (OU LES) GT LE PLUS RECENT, POSTERIEUR A LA FOURCHETTE CHRONOLOGIQUE (en NMI)

GT	Nbre	data inf	data sup
to11d	1	1433	1568
to9b	1	1435	1545

Matériel intrusif/à la fourchette chronologique estimée

effectif en NMI 2

% / à l'effectif total des NMI 2,3

Matériel redéposé/à la fourchette chronologique estimée (que sur redep ; pas GR)

effectif en NMI 3 % / à l'effectif total des NMI 3,5

Totalité du matériel redéposé en NMI (redep +GR)

effectif en NMI 3 % / à l'effectif total des NMI 3,5

Totalité du matériel redéposé en NR (redep + GR)

effectif en NR 0 % / à l'effectif total 0

Pour les médiévistes : matériel redéposé Gallo-Romain (en NMI)

effectif en NMI 0 % / à l'effectif total des NMI 0

Pour les médiévistes : matériel redéposé Gallo-Romain (en NR)

effectif en NR 0 % / à l'effectif total 0

Fig.6 Help to date a domestic dump (Set 28) of the castle of Tours

3.3 Reliability in archaeological publications

Jean-Claude Gardin's logicism aim to publish the results with the administration of evidence [8, 9], Rigny's forthcoming electronic publication is based on this formal structure [4] when the Dataset of this excavation is a part of ArSol Database. The reader can thus choose between traditional reading and in-depth consultation. For each proposal, the reader can go back to the basis of the proposals, from the most synthetic proposals to the evidence mobilized (in particular, field observations). Like all model, logicist structuring is a reduction, without the rhetorical apparatus traditionally used in publications, but it preserves all the constitutive elements of the cognitive construction. So, this form of writing contributes to reduce the imbalance observed between bibliographic production and our consuming capacities. Rigny publication opens the way to a new form of publication adapted to the growing preponderance of consultation over reading. This is the first publication of an excavation

monograph in logicist writing, but there are several publications in archaeology of techniques available on Arkeotek Journal (http://www.thearkeotekjournal.org/).

The purpose was to make explicit the steps of the reasoning by distinguishing the basic data (or "initial propositions"), and the inference operations carried out on these data to establish the interpretative hypotheses (Fig.7). Therefore it is possible to constitute a tree structure which gives a synoptic representation of the argumentation and enables a quick assessment of its relevance. The argumentation takes the form of a series of inference operations from {P0} (initial propositions) to {Pn} (final propositions) via intermediate propositions {Pi} [9, 17].

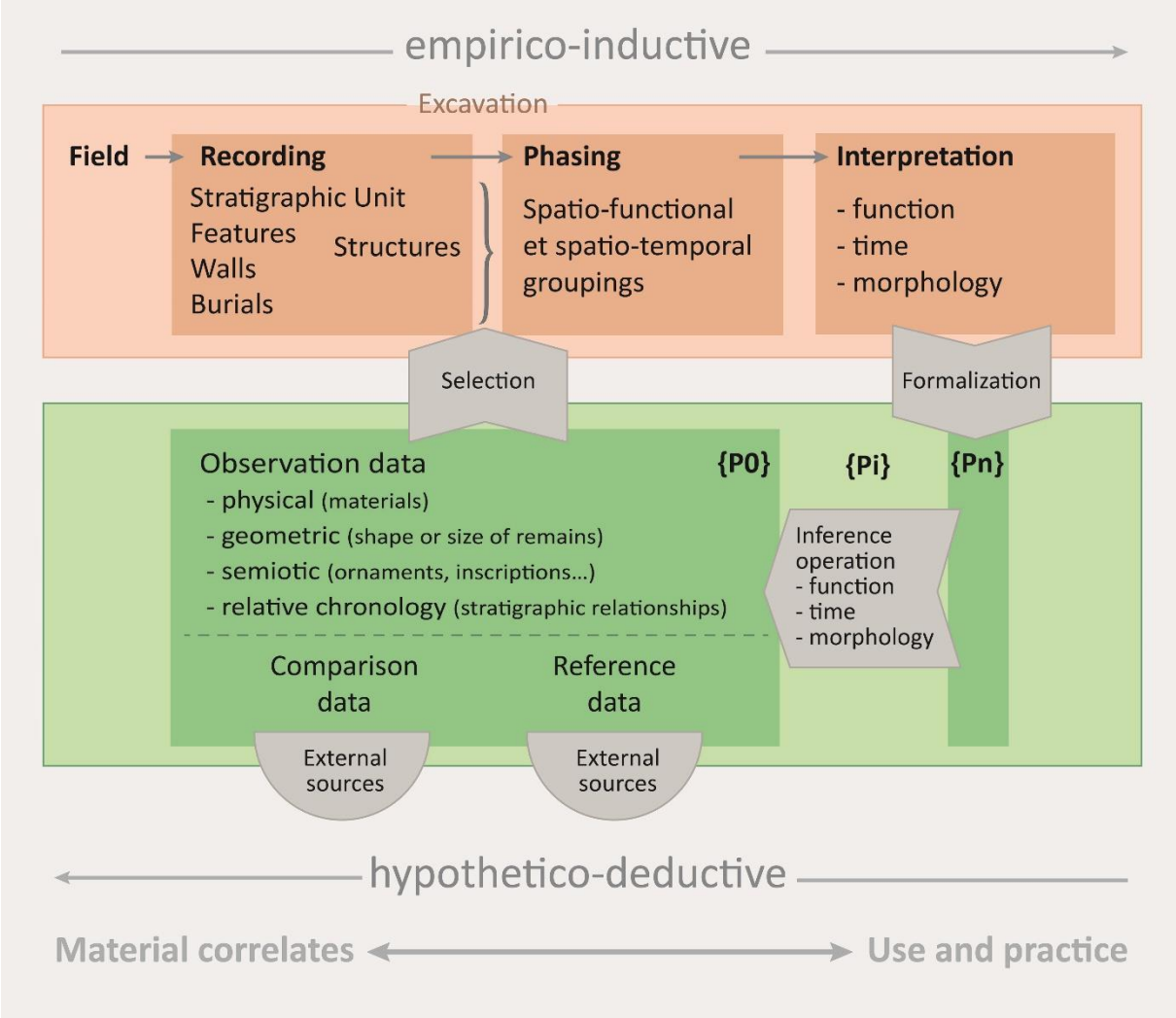


Fig.7 Diagram showing the structuring of logicist writing

The use of logicist analysis leads to a reduction of the published text. If we compare this publication to a classic monograph of excavation, readers have access to the whole reasoning and arguments. Logicist writing highlights the chains of inference bridging the gap between empirical facts (or descriptive propositions) at one end of the argumentation, and final propositions (or conclusions) at the other end. The highlighting of the interpretative processes allows their critical assessment. Moreover, this very structured writing opens the way for different levels of reading, from speed-reading of the results to the in-depth examination of the evidence. In Rigny's publication, hyperlinks provide direct access to field recording data in ArSol database. You can see an example here with the M42 wall.

Another way to consult Rigny publication is to browse within the logicist diagrams. Logicist diagrams display the argumentation in the form of an inference tree developing from left to right. These diagrams provide a graphic overview of the argumentation (Fig.8). They are interactive and allow access to the detailed argumentation. The diagrams are automatically produced through the XML TEI encoding of the texts [4].

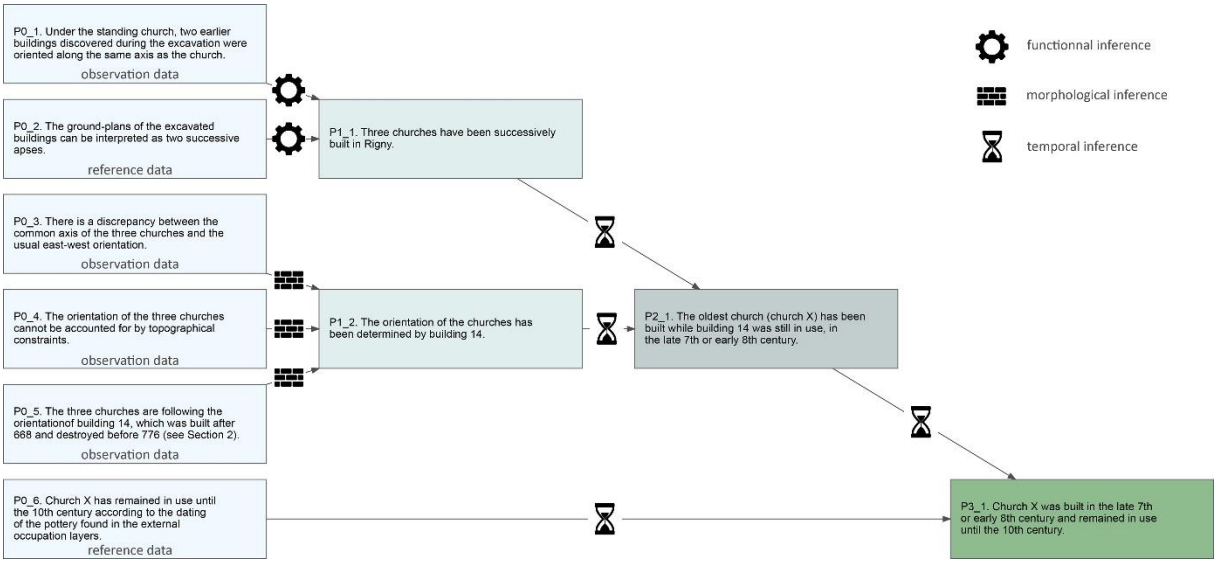


Fig.8 Example of logicist diagram for Rigny publication with typing of inferences

As with ArSol data, the Rigny publication data (a TEI file) has been mapped to CIDOC CRM to allow interoperability. The proposals and inference links formalized in XML-TEI are well suited for mapping with CIDOC, particularly by using the CRMinf extension [18]. For example, it was possible to model the inference types used in the reasoning (functional, morphological or temporal inferences) and to enter them in the TEI file. In addition, thus modelled, the reasonings themselves can constitute knowledge bases that will feed the Linked Open Data [15], but it would need more of this kind of online datasets to have the benefit of bringing the reasonings themselves to the Linked Open Data.

Conclusion

By complying with FAIR principles, data can be reused, but the reusability of such data is only meaningful if it can be ensured that it is relevant and reliable. Allowing data traceability on several aspects is already a good approach. ArSol aims to enable this traceability by offering a scan of the initial field record associated with the computerized data, by giving the possibility to consult the statistical methods implemented to date an archaeological context and by allowing a scientific publication to refer to field data as evidence in a support of an argumentation.

From the field recording during the excavation to the reasoning used in the publication, the ArSol database is the backbone of a robust system. Although there is still much to be done to improve the reliability of our data, the ArSol system provides the possibility to trace data from acquisition to publication.

Modelling this traceability would be an interesting task.

Acknowledgments

We would like to thank the MASA Consortium of the TGIR Huma-Num and the H2020 ARIADNEplus program, a project funded by the European Commission under the H2020 Programme, contract no. H2020-INFRAIA-2018-1-823914.

Bibliography

1. Bellanger L, Husi Ph (2012) Statistical tool for dating and interpreting archaeological contexts using pottery, Journal of Archaeological Science. Volume 39, Issue 4, April 2012, pp. 777–790. http://dx.doi.org.gate3.inist.fr/10.1016/j.jas.2011.06.031 Accessed 25 January 2019

2. Bellanger L, Husi Ph, Tomassone R (2006) Statistical aspects of pottery quantification for dating some archaeological contexts, Archaeometry, n°48, Oxford, pp. 169-183.

3. Berners-Lee Tim (2006). Linked Data. https://www.w3.org/DesignIssues/LinkedData Accessed 25 January 2019

4. Buard P-Y, Zadora-Rio E (2019) Publishing an Archaeological Excavation Report in a Logicist Workflow. In: Open Open Open. The Rise of Open Scientific Publishing and the Archaeological Discipline: Managing the paradigm Shift, Proceedings of the 46rd Annual Conference on Computer Applications and Quantitative Methods in Archaeology, Archaeopress, Oxford.

5. FAIR principles, FORCE11 https://www.force11.org/group/fairgroup/fairprinciples Accessed 25 January 2019

6. Galinié H (2013) Les clefs du sol (1976-77), Les petits cahiers d'Anatole 25, novembre 2013. http://citeres.univ-tours.fr/doc/lat/pecada/pecada_25.pdf Accessed 25 January 2019

7. Galinié H, Husi Ph, Rodier X, Theureau Ch, Zadora-Rio E (2005) ARSOL. La chaine de gestion des données de fouilles du Laboratoire Archéologie et Territoires*, Les petits cahiers d'Anatole* 17, mai 2005. http://citeres.univ-tours.fr/doc/lat/pecada/F2_17.pdf Accessed 25 January 2019

8. Gardin J-C (1980) Archaeological Constructs: An Aspect of Theoretical Archaeology, Cambridge University Press, Cambridge.

9. Gardin J-C, Guillaume O, Herman P Q, Hesnard A, Lagrange M-S, Renaud M, Zadora-Rio E (1987) Systèmes experts et sciences humaines : le cas de l'archéologie*,* Eyrolles, Paris.

10. Holmen J, Ore C-E S, Eide Ø (2004). Documenting two histories at once: Digging into archaeology, Oxford. ISSN 0143-3059. 1227, pp. 221-224. http://www.edd.uio.no/artiklar/arkeologi/documenting%20two%20histories.html Accessed 17 june 2019

11. Husi Ph, Rodier X (2011) ArSol: an archaeological data processing system. In: Erzsébet J., Redő F., Szeverényi V. (dir.) On the Road to Reconstructing the Past, Proceedings of the 36th CAA Conference, Budapest, 2-6 April 2008, pp. 86-92. https://proceedings.caaconference.org/files/2008/CD36_Husi_Rodier_CAA2008.pdf Accessed 25 January 2019

12. ICERAMM: Information sur la Céramique Médiévale et Moderne http://iceramm.univ-tours.fr/ Accessed 25 January 2019

13. Linked Data Finland, Living Laboratory Data Service for the Semantic Web http://www.ldf.fi/ Accessed 17 June 2019

14. Marlet O, Curet S, Rodier X, Bouchou-Markhoff B (2016) Using CIDOC CRM for dynamically querying ArSol, a relational database, from the semantic web. In: Stefano Campana, Roberto Scopigno, Gabriella Carpentiero, Marianna Cirillo. CAA2015 Keep the revolution going, proceedings of the 43rd Annual CAA Conference, Sienna, 30 March-3 April 2015, Archaeopress Archaeology, 2016, pp. 241-250.

15. Marlet O et al. (2019) The Archaeological Excavation Report of Rigny: An Example of an Interoperable Logicist Publication, Heritage 2019, 2(1), pp. 761-773 https://doi.org/10.3390/heritage2010049 Accessed 17 June 2019

16. Marlet O., Francart T., Markhoff B, Rodier X (2019) OpenArchaeo for Usable Semantic Interoperability, in : Poggi A. (ed.) Proceedings of the First International Workshop on Open Data and Ontologies for Cultural Heritage (Rome, Italy, June 3, 2019), Central Europe Workshop Proceedings, vol. 2375, pp. 5-14. http://ceur-ws.org/Vol-2375/paper1.pdf Accessed 17 June 2019

17. Roux V, Aussenac-Gilles N (2013) Knowledge Basis and Query Tools for a Better Cumulativity in the Field of Archaeology: The Arkeotek Project. In: CAA 2011, Fusion of Cultures. Proceedings of the 38th Annual Conference on Computer Applications and Quantitative Methods in Archaeology, Granada, Spain, April 2010, Archaeopress, 2013, pp. 267-272.

18. Stead S, Doerr M (2015) CRMinf: the Argumentation Model. An Extension of CIDOC-CRM to support argumentation, Version 0.7 https://www.ics.forth.gr/isl/CRMext/CRMinf/docs/CRMinf-0.7.pdf Accessed 17 June 2019