Modéliser l'information archéologique à l'ère du Web sémantique

Muriel Van Ruymbeke

Dès l'Antiquité, et peut-être même avant, des individus se sont intéressés au passé. Qu'il soit visible, au travers de traces laissées dans le paysage, ou invisible mais évoqué dans les souvenirs et autres récits, ce monde évanoui a sollicité les esprits curieux. Il ne s'agissait pas encore d'une science, au sens moderne du terme, avec ses protocoles et ses codes. Lorsque Platon, vieillissant, vante l'*ἀρχαιολογία*, il veut donner à voir que la connaissance du passé est une aide précieuse pour comprendre le présent et bâtir l'avenir idéal.

Bien sûr, nous savons que le but de l'archéologie au sens où l'entendait Platon ne visait pas à reconstituer un passé le plus probable possible mais plutôt à participer au monde en devenir et peu importe si cela supposait de prendre quelques libertés avec la réalité. Nous n'en sommes plus là et aujourd'hui l'archéologie est bien une discipline scientifique, composée de concepts, de méthodes, de postures et des inévitables conflits épistémologiques qui les accompagnent. En effet, les pratiques ont beaucoup évolué. Elles ont obéi à des modes, des approches. L'approche historico-culturelle a par exemple été opposée à l'archéologie processuelle, elle-même contestée par la suite par l'approche post-processuelle.

Par-delà ces différences de faire de l'archéologie, la manière de l'exprimer a pourtant peu varié. Du carnet de fouille à la publication définitive en passant par le rapport de fouille ou les études synthétiques, analytiques, thématiques, les données archéologiques ont longtemps été élaborées et transmises essentiellement par écrit. L'usage de l'informatique a élargi les perspectives scientifiques et désormais l'information archéologique se crée et se transmet différemment. Les données sont à présent modélisées et/ou numérisées. Aujourd'hui, l'archéologie est entrée dans une nouvelle ère, celle de l'archéologie digitale.

L'usage omniprésent de l'informatique se manifeste principalement dans la manière dont on enregistre les données observées et celle dont on les transmet. Au-delà d'un indéniable apport méthodologique, les nouvelles pratiques génèrent cependant de nouveaux écueils. En effet, pour pouvoir profiter de la puissance de l'ordinateur, il est nécessaire de modéliser les données qu'on lui soumet. Mal conduite, cette démarche peut aboutir à un appauvrissement de l'information. En archéologie, le processus de modélisation susceptible d'éviter ce danger ne fait pas encore consensus.

La recherche doctorale qui est présentée ici a consisté à concevoir, tester et finalement proposer de nouvelles pistes de modélisations pour les données archéologiques. Ces propositions tiennent compte des spécificités des données archéologiques. Elles ont également l'avantage de se conformer aux normes du Web sémantique et de son corollaire, les *Linked Open Data*. Leur solidité a été éprouvée en les appliquant

à un corpus riche et consistant, celui des données archéologiques antiques et alto-médiévales de la commune de Theux.

Modélisation

Si l'on observe que les verbes digitaliser, numériser, modéliser, sont régulièrement employés indifféremment l'un pour l'autre, les activités qu'ils désignent sont pourtant à distinguer. On peut certes concéder que les verbes digitaliser et numériser recouvrent pratiquement le même sens, c'est-à-dire que les données initiales sont transformées en une suite de caractères et de nombres. Il n'en va cependant pas de même pour le verbe modéliser.

Dans l'ouvrage qu'il a consacré à la modélisation de l'information en archéologie et en anthropologie, Cesar González-Pérez soutient que modéliser consiste à créer avant tout une représentation mentale de quelque chose : l'horaire de la journée à venir, la liste des courses à faire ou le schéma d'une organisation. Selon C. González-Pérez toujours, un modèle ne l'est que s'il possède trois caractéristiques : s'il représente un objet, s'il le simplifie et s'il permet de raisonner à son propos. En ce sens, les cartes de distribution de sites ou d'artefacts archéologiques sont des modélisations ; les matrices de Harris aussi.

Dans tous les cas de figure, l'opération de modélisation transforme la représentation de l'information originale dans le but d'en obtenir plusieurs avantages. Ceux-ci sont aussi divers et variés que le sont les méthodologies de modélisation. Ainsi, un inventaire établi sous la forme d'une base de données informatique sera beaucoup plus ergonomique à exploiter que le même inventaire établi sur des fiches en papier : en fonction des champs indexés, on pourra le questionner sous de multiples angles, lancer des requêtes combinées et dresser des statistiques.

Le Web sémantique et les données liées (*Linked Open Data*)

La notion de Web sémantique a été conçue et expliquée, à l'entame du 21e siècle, par Sir Tim Berners-Lee. Cet informaticien britannique également réputé pour être l'un des créateurs, avec notre compatriote Robert Cailliau, du *World Wide Web* a démontré en 2001 l'intérêt qu'il y aurait, dans un monde interconnecté, à décloisonner l'information. En effet, bien que les bases de données fussent facilement accessibles grâce à Internet, leur gestion indépendante les isolait les unes des autres. La proposition de T. Berners-Lee reposait sur l'idée que, plutôt que d'enregistrer à de multiples reprises les mêmes éléments dans des bases de données différentes, il était plus intéressant d'identifier les concepts redondants et de ne les enregistrer qu'une seule fois. Restait à imaginer et développer un moyen de lier les données entre elles lorsqu'elles étaient en interaction. C'est le concept de données liées, intimement associé à celui du Web sémantique.

En vingt ans, ces propositions ont été adoptées et surtout concrétisées. On observe aujourd'hui une vaste trame de données suffisamment interconnectées pour permettre leur partage, leur échange et, encore mieux, leur exploitation au profit de la découverte de nouvelles connaissances. Ces données liées et ouvertes sont gratuitement accessibles sur le Web, lisibles par les ordinateurs, compréhensibles pour les humains et sémantiquement assemblées. Elles respectent les normes du *World Wide Web Consortium* (W3C) dont en particulier, mais pas uniquement, celles de la représentation RDF (*Resource Description Framework*). Ces normes permettent la création, la publication et la mise en relation de métadonnées et de systèmes d'organisation des connaissances de telle sorte que la signification des termes soit claire à la fois pour les humains mais aussi pour les ordinateurs. Pour décrire la chose brièvement, le Web sémantique repose sur

un tissu de données accessibles à tous (humains et machines) et interopérables. Il permet toutes les requêtes et, de ce fait, favorise la création de nouvelles connaissances.

Modélisation sémantique en archéologie

Cette nouvelle manière de penser et d'utiliser les bases de données et les ressources du Web a rapidement intéressé les professionnels du patrimoine culturel en général. Par exemple, en 2013, Patrick Le Boeuf a écrit à propos des inventaires de musée : « nous attendons autre chose qu'une simple juxtaposition d'inventaires en ligne isolés les uns des autres. Nous souhaitons pouvoir aussi consulter plusieurs inventaires simultanément, resituer un objet patrimonial dans un contexte en voyant les autres objets avec lesquels il est en relation à un titre ou à un autre... » (Le Bœuf, 2013 : 1). Au-delà du monde des catalogues de musées ou de bibliothèques, cette ambition vaut également pour le domaine de l'archéologie qui a grand intérêt à décloisonner les savoirs, les mettre en relation les uns avec les autres, les réutiliser et les faire fructifier.

Pourtant, aujourd'hui, dans le domaine du patrimoine culturel, seuls quelques projets fonctionnent effectivement sur le principe des données ouvertes et liées. Il s'agit de rares catalogues de bibliothèques comme celui de la BNF ou de quelques catalogues de musées comme ceux du British ou du Getty Museum par exemple. On peut également citer les agrégateurs d'objets culturels numérisés comme le méta catalogue européen Europeanna.

Ce faible succès s'explique en grande partie par les difficultés rencontrées pour modéliser les données archéologiques d'une part et pour respecter les normes du Web sémantique d'autre part. La difficulté de modéliser les données archéologiques tient à leur hétérogénéité, que l'on peut décrire comme un agglomérat de composantes spatiale, temporelle et fonctionnelle. La difficulté réside également dans leur imperfection : les données sont souvent incertaines, incomplètes, imprécises, multiples voire contradictoires. L'accord avec les normes du Web sémantique se heurte aux difficultés suivantes : l'archéologie ne connaît pas encore une vraie standardisation. Les vocabulaires et les protocoles utilisés ne font pas consensus au-delà d'un cercle restreint d'utilisateurs. Pire, la notion de données ouvertes et accessibles est loin d'être communément acceptée.

Le modèle MIDM

La première étape de la recherche présentée ici a donné lieu à la création du modèle MIDM (*Mutiple Interpretation Data Model*). Construit pour être publié au sein de la communauté archéologique internationale, il a par conséquent dès l'origine été composé en anglais. Cette langue est conservée ici afin d'éviter toute ambiguïté relative aux éléments décrits ou illustrés.

La réflexion a démarré à partir de l'hypothèse suivante : si l'on veut obtenir une modélisation des données archéologiques qui soit viable, il faut pouvoir clairement différencier les faits archéologiques des discours tenus à leur sujet. Il est également nécessaire que le modèle supporte les imperfections des données. Une première phase de conceptualisation a été exécutée. Elle a ensuite été formalisée en UML pour donner le modèle conceptuel MIDM. Il est proposé ci-dessous sous la forme d'un diagramme de classes simplifié, c'est-à-dire présenté sans les attributs. Il est formé de classes liées entre elles soit par des relations structurelles dotées de cardinalités, soit par une dépendance hiérarchique, appelée aussi relation de généralisation. Cet ensemble décrit de manière théorique les éléments constitutifs de l'information archéologique au sens large ainsi que leurs interactions.

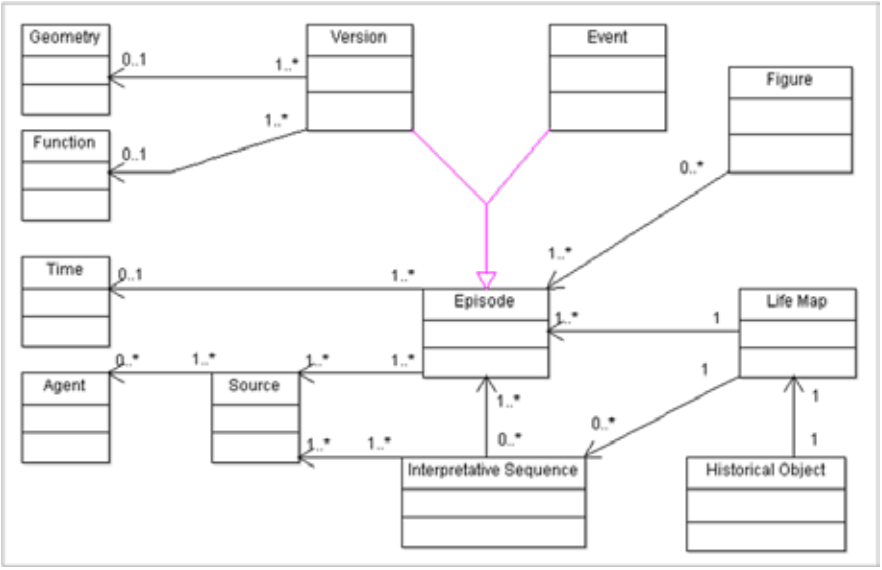

Fig. 1

Le modèle MIDM reprenant en noir les relations structurelles et en fuchsia les relations de généralisation © M. Van Ruymbeke

Le modèle MIDM contient douze classes. La classe de départ est la classe intitulée *Historical Object*. Cet objet peut se définir comme « tout concept ou élément physique, animé ou non, existant, ayant existé ou réputé tel, qui peut éventuellement être composé d'états successifs ». Cet objet, de même que son contexte de vie ne peuvent plus être observés directement de sorte que leurs spatialités, leurs temporalités et leurs fonctions ne peuvent être approchés qu'au moyen d'hypothèses.

Toutes les autres classes du modèle font référence aux hypothèses se référant à l'objet historique. Ces hypothèses concernent la localisation et la géométrie (*Geometry*), la datation (*Time*), la fonction (*Function*), et les sources documentaires (*Source*) qui les étayent. Ces hypothèses peuvent concerner un état de l'objet et l'on parlera d'une *Version* de l'objet, ou un événement (*Event*) l'ayant impacté. Ces deux classes sont les spécialisations d'une classe plus générale appelée *Episode*. Un épisode peut mettre en scène une ou plusieurs figure(s) historique(s) (*Figure*). Les hypothèses peuvent concerner l'évolution d'un objet et l'on parlera de séquence interprétative (*Interpretative Sequence*) Ces hypothèses ont un ou plusieurs auteur(s) (*Agent*), également présent(s) dans le modèle. Enfin, la carte de vie (*Life Map*) est une classe qui désigne l'ensemble des hypothèses émises à propos d'un objet historique.

L'originalité du modèle MIDM réside dans la création du concept d'épisode, mais également (et même surtout) dans la création du concept de séquence interprétative. Ensemble, ils permettent de bien distinguer les reconstructions interprétatives de la réalité qu'ils décrivent. Le jeu des relations entre ces concepts et leurs cardinalités permet de travailler avec les imperfections des données

archéologiques ainsi qu'avec le maniement d'interprétations multiples. En cela le modèle MIDM, tout en adoptant un certain nombre des propriétés de l'Objet historique tel que défini dans d'autres modèles (comme par exemple dans le modèle OHFET), s'en distancie au niveau de la définition de son identité, puisque le modèle MIDM admet que cet objet change au fil du temps.

Le modèle apparié avec l'ontologie du CIDOC CRM

En principe, un modèle conceptuel n'a d'utilité que s'il peut être implémenté. Dans le cas présent, cette étape a exigé de poser un choix entre l'implémentation sous forme de base de données relationnelle (SQL) ou l'implémentation dans une base de données orientée graphe (*graph-oriented* NoSQL). En tenant non seulement compte de l'émergence scientifique et technique du Web sémantique et des ontologies standardisées, mais également de l'efficacité des bases de données orientées graphe en matière d'exploitation des relations, il a semblé plus judicieux de matérialiser le MIDM en optant pour la deuxième option.

Le fait de travailler avec les ontologies consiste à stocker l'information non pas dans des bases de données comportant des champs liés entre eux, mais dans un système de graphes dans lesquels l'information est décomposée en triplets. Ces triplets sont des petites phrases composées d'un sujet, d'un verbe et d'un objet. Les sujets et les objets sont classés dans des catégories de concepts organisés hiérarchiquement. Les verbes sont classés dans des catégories de propriétés, elles aussi organisées hiérarchiquement, et lient les concepts entre eux.

Restait à sélectionner un langage de métadonnées dédié à la représentation des données culturelles ou patrimoniales. Parmi les rares ontologies culturelles disponibles, le CIDOC CRM et ses extensions compatibles ont été considérés comme le meilleur choix et ce, pour plusieurs raisons. Tout d'abord, il s'agit de l'ontologie la plus complète en matière de patrimoine puisque si, au départ, elle ciblait presque exclusivement l'information muséale, elle a rapidement étendu son périmètre au patrimoine archéologique, enfoui ou bâti ainsi qu'à la documentation bibliographique.

D'un autre côté, le CIDOC CRM est devenu en 2006 une norme ISO reconnue. Cette norme a été actualisée en 2014 et est actuellement à nouveau en phase de révision. Outre ces révisions officielles, la norme est en réalité amendée et complétée environ tous les trois mois. Le CIDOC CRM est nourri et validé par un grand nombre de chercheurs, mais surtout il est partagé par un assortiment croissant d'utilisateurs. Un certain nombre d'établissements culturels français travaillent à apparier leurs ressources avec les concepts du CIDOC, c'est le cas par exemple du projet Hadoc ou de certaines bases de données archéologiques. Le consortium MASA organise des cursus pour former les archéologues français à son utilisation. Cette popularité, outre son aspect rassurant, procure un argument intéressant en matière d'interopérabilité et d'intégration ultérieures.

Il faut bien reconnaître que l'appariement des concepts définis dans le modèle MIDM avec ceux définis par les ontologies du CIDOC CRM a présenté un certain nombre d'obstacles. L'un d'entre eux, par exemple, a été l'instabilité des ontologies CIDOC CRM. En effet, leurs concepteurs les modifient et les mettent à jour plusieurs fois par an. Comme les modifications d'une version à l'autre sont parfois importantes, il a été nécessaire d'adapter les appariements au fur et à mesure de ces modifications. D'un autre côté, le manque de profondeur et de subtilité des *thesauri* actuels ne permet pas d'exprimer toute la sémantique fonctionnelle associée à un objet avec les nuances désirées. Il faut à ce sujet saluer l'initiative HyperThesau qui vise, notamment, à remédier à cette faiblesse.

Malgré l'abondance de classes et propriétés spatiotemporelles disponibles dans le CIDOC CRM, ce sont les ontologies bien connues GeoSPARQL et Owl-Time qui ont été privilégiées pour l'appariement des concepts du MIDM de *Geometry* et *Time*. En effet, ces ontologies peuvent être associées à des traitements de raisonnement qui autorisent les requêtes spatiales, temporelles et donc spatio-temporelles.

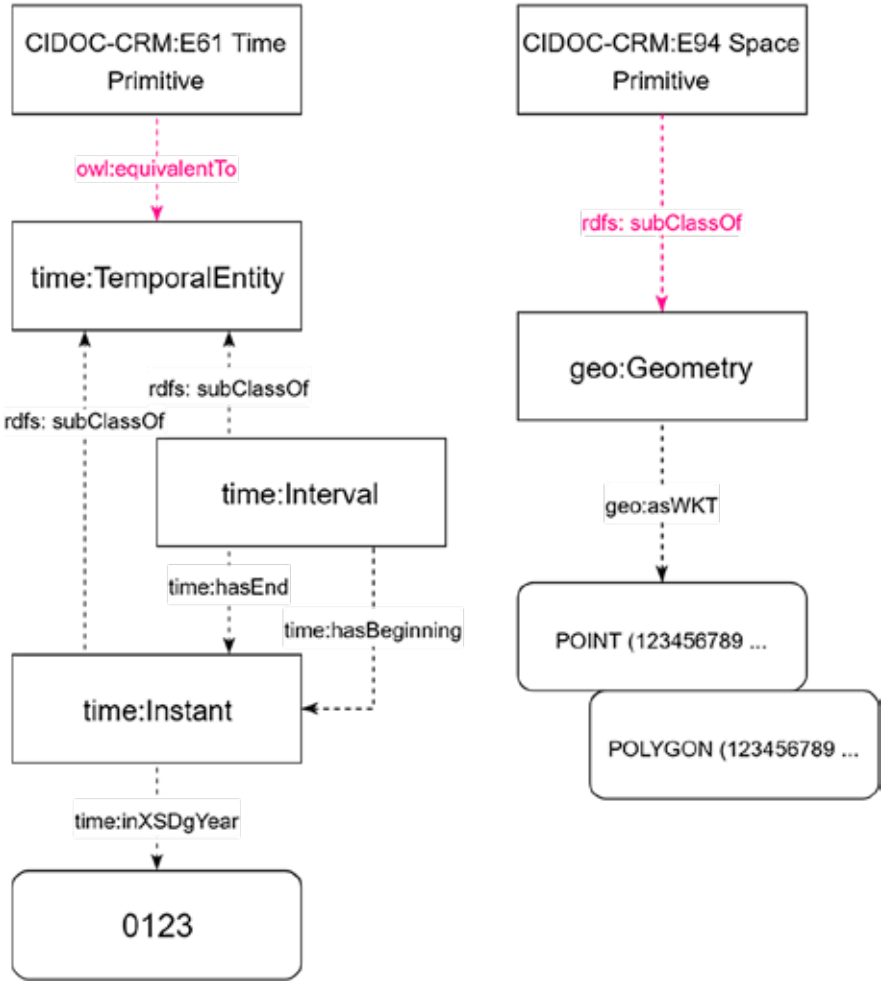

Fig. 2
Liens créés entre le CIDOC CRM, OWL-Time et GeoSPARQL © M. Van Ruymbeke

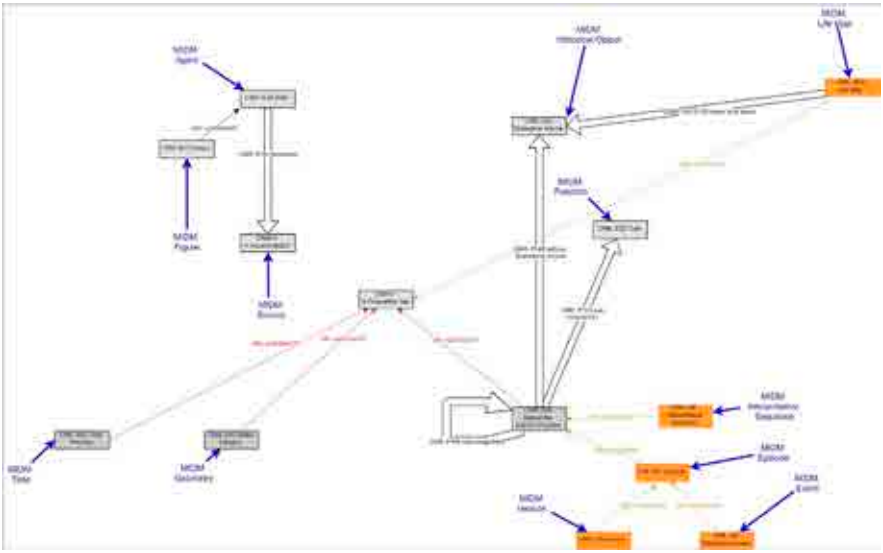

Fig. 3
Appariement des concepts du MIDM (en bleu) vers les classes et propriétés des ontologies du CIDOC (en blanc et gris) selon la cinquième piste d'appariement (les nouvelles classes sont en orange). Les éléments en rose représentent les propositions de modification de dépendance hiérarchiques du CIDOC – CRM 6.2.6. © M. Van Ruymbeke

L'appariement des concepts du MIDM a été réalisé et même en plusieurs exemplaires puisque, compte tenu des changements survenus dans les ontologies du CIDOC, ce ne furent pas moins de cinq pistes d'appariement qui ont été tracées. Chacune d'entre elles a abouti à la création de nouvelles classes et propriétés, afin d'exprimer pleinement les notions d'épisodes, versions et séquences définis dans le MIDM. C'est la dernière des pistes expérimentées qui a été utilisée pour l'instanciation du cas d'étude : les données archéologiques antiques et alto-médiévales de la commune de Theux.

La preuve conceptuelle au moyen des données antiques et alto-médiévales de la commune de Theux

Cette partie du travail a consisté à tester le modèle MIDM implémenté sous forme d'ontologie en le peuplant de données. Délivrant des antiquités depuis le 16e siècle, fouillée depuis le milieu du 19e siècle, la commune de Theux se présentait comme un cas d'étude idéal.

Les données archéologiques recueillies sont diverses et variables à tous points de vue. Ce sont des mentions de découvertes fortuites, des rapports de fouilles succincts, des listes d'objets, des lettres manuscrites, mais également des rapports de fouilles détaillés et

une monographie de site. Ces centaines de références n'atteignent pas toutes le même degré d'exactitude ou de précision et leurs affirmations ne sont, en général, plus vérifiables. Elles constituent autant d'hypothèses et d'interprétations se superposant au fil du temps.

La collation et la relecture de cette somme de données a été traitée de la manière suivante : toutes les hypothèses de localisations de tous les sites antiques et alto-médiévaux de la commune de Theux ont été vectorisées et géoréférencées sous forme de points, de lignes ou de polygones dans le logiciel ArcGIS Pro puis transformées en format WKT afin de peupler la base de connaissances. Les hypothèses de restitutions chronologiques ont également été codées en suivant le protocole requis. Les données fonctionnelles ont été exprimées à l'aide des concepts des *thesauri* PACTOLS et PATRIARCHE.

L'ensemble des données archéologiques antiques et alto-médiévales de la commune de Theux, en l'état des connaissances daté de septembre 2020, a été exprimé sous la forme de triplets. Cette phase d'instanciation fut extrêmement longue et fastidieuse à réaliser. Elle était cependant nécessaire pour valider les étapes de conceptualisation précédentes. Elle a permis de vérifier que la piste d'appariement sélectionnée permettait d'exprimer valablement des données réelles. Elle a également rendu possible la mise en évidence de la pertinence du modèle MIDM. Enfin, elle a démontré que l'on pouvait exercer des requêtes spatio-temporelles mais également sémantiques sur l'ensemble d'une information archéologique représentée sous forme d'un graphe de données.

Conclusion

Face aux constantes mutations de notre environnement technologique, il est important d'adapter nos pratiques professionnelles. Il n'est pas un archéologue qui dira le contraire. Cependant, cette adaptation ne peut se réussir que si elle tient pleinement compte des fondements épistémologiques de notre discipline. Ces piliers scientifiques garantissent la durabilité mais également la valeur de l'archéologie. La recherche brièvement résumée ici a consisté à proposer une approche de modélisation respectueuse de tout ce qui constitue l'archéologie contemporaine. Cette approche a également veillé à tenir compte des besoins à venir principalement en matière d'ouverture et d'interopérabilité des données. Si l'exercice peut paraître extravagant voire intrépide, il doit être perçu comme une invitation à poursuivre activement les recherches et les développements dans le domaine des données ouvertes et liées (*Linked Open Data*) en archéologie.

Merci à Gilles-Antoine Nys pour la relecture attentive et scientifique de cet article.

Bibliographie

Berners-Lee T. et Hendler J., 2011. Publishing on the semantic web. *Nature*, 410, 6832 : 1023-1104.

Berners-Lee T., Hendler J. et Lassila O., 2011. The Semantic Web. *Scientific American*, mai 2001, vol. 284, 5 : 3443.

Boissinot P., 2015. *Qu'est-ce qu'un fait archéologique ?* Paris, École des Hautes Études en Sciences sociales.

De Runz C., 2008. *Imperfection, temps et espace : modélisation, analyse et visualisation dans un SIG archéologique*. Reims, Université de Reims, Champagne-Ardenne.

Djindjian F., 2017. *L'archéologie - Théorie, méthodes et reconstitutions*. Malakoff, Armand Colin (2e éd.).

Doerr M., Bekiari C., Bruseker G., Ore C.-E., Stead S. et Velios T., 2021. *Definition of the CIDOC Conceptual Reference Model v7.1.1 (Version v7.1.1)*. The CIDOC Conceptual Reference

Model Special Interest Group. URL https://doi.org/10.26225/FDZH-X261

Fusco G., Bertoncello F., Candau J., Emsellem K., Huet T., Longhi C., Poinat S., Primon J.-L. et Rinaudo C., 2014. *Faire science avec l'incertitude : réflexions sur la production des connaissances en Sciences Humaines et Sociales*. Nice, Open access Halshs-01166287.

Gelé A., 2014. Objet externalisé et objet vecteur de sens. De l'archéologie des périodes modernes à l'archéologie historique, un état de la question. *Europa Moderna. Revue d'histoire et d'iconologie*, vol. 4, 1 : 420.

Geser G., 2016. *Towards a Web of Archaeological Linked Open Data*. Ariadne.

González-Pérez C., *Information Modelling for Archaeology and Anthropology*, Cham, Springer International Publishing, 2018.

González-Pérez C., 2018. *Information Modelling for Archaeology and Anthropology*. Cham, Springer International Publishing.

Juanals B. et Minel J.-L., 2020. Stratégies éditoriales des musées. Une approche de la médiation par l'accès ouvert aux données numérisées. *Culture & Musées. Muséologie et recherches sur la culture*, juin, 35 : 4975.

Le Boeuf P., 2013. De la sémantique des inventaires aux musées en dialogue : la modélisation CIDOC CRM. *Culture & musées*, 22 : 89111.

Martin-Rodilla P. et González-Pérez C., 2018. *Representing Imprecise and Uncertain Knowledge in Digital Humanities: A Theoretical Framework and ConML Implementation with a Real Case Study*. Salamanca, ACM Press.

Migliorini S., 2018. Enhancing CIDOC-CRM models for GeoSPARQL processing with mapreduce. *Workshop Proceedings, 2230 (2018); 2nd Workshop On Computing Techniques For Spatio-Temporal Data in Archaeology And Cultural Heritage, Melbourne, Australia [AU]*, 28 août 2018, 2230 : 5165.

Nouvel B., 2019. Le thésaurus PACTOLS, système de vocabulaire contrôlé et partagé pour l'archéologie. *Archéologies numériques*, 3, 1.

Nys G.-A., Van Ruymbeke M. et Billen R., 2018. Spatio-temporal reasoning in CIDOC CRM : an hybrid ontology with GeoSPARQL and OWL-Time. *CEUR Workshop Proceedings*, 13 octobre 2018, 2230 : 37-50.

Perrin E., 2021. Thésaurus et interopérabilité des données archéologiques : le projet HyperThesau. *Humanités numériques*, 2021, 4.

Perry S. et Taylor J.S., 2018. *Theorising the Digital. A call to Action for the Archaeological Community. Oxford*, Archaeopress.

Van Ruymbeke M., 2021. *Modéliser l'information archéologique à l'ère du web sémantique - Relecture 2.0 des données archéologiques antiques et alto-médiévales de la commune de Theux (B.)*, Université de Liège, thèse de doctorat non publiée.

Van Ruymbeke M., Carré C. et Billen R., 2012. L'existant et l'ayant existé. Documenter le patrimoine dans la diachronie. *Thema & Collecta*, 2 : 4251.

Vincent M.L., Levy T.E., Bendicho V.M.L.-M. et Ioannides M., 2017. *Heritage and Archaeology in the Digital Age: Acquisition, Curation, and Dissemination of Spatial Cultural Heritage Data*. Springer.

Weil R., 1959. *L'« archéologie » de Platon*. Paris, C. Klincksieck.