Le CIDOC-CRM : principes de base et outil en ligne disponible pour évaluer la conformité des systèmes d'enregistrement de terrain avec un modèle conceptuel cible

discussion méthodologique – discussion de projet

- projet professionnel au sein de l'INRAP ;

- thématiques de l'atelier concernées : 1 - méthodes et outils d'enregistrement et de traitement des données de terrain ;

présentation

Christophe Tufféry[1] (intervention préparée par Christophe Tufféry et Emeline Le Goff[2])

1. Rappel général sur la norme CIDOC-CRM et l'extension CIDOC-CRMarchaeo

Le CIDOC-CRM est un modèle conceptuel de référence (ou ontologie) de haut niveau pour décrire les données du domaine culturel. Le numéro de version actuel de la norme CIDOC-CRM est la version 5.0.4 publié comme norme ISO 21127:2014. Plusieurs extensions du CIDOC-CRM existent. Elles élargissent le domaine du CIDOC-CRM dans des domaines spécialisés[3].

Parmi ces extensions, CRMarchaeo est une extension spécialisée du domaine de l'archéologie, développée depuis 2013 par l'Institute of Computer Science de la Fondation Forth (Fondation pour la Recherche et la Technologie) en Grèce (Doerr M. et al. 2017). CRMarchaeo est une ontologie qui permet de désigner les entités d'une opération archéologique (prospection, diagnostic, fouille, etc.) et leurs relations ainsi que la documentation produite et le mobilier archéologique découvert.

L'objectif de cette extension est de fournir les moyens de documenter les fouilles pour permettre de :

– maximiser la capacité d'interprétation pendant ou après une opération ;

– formaliser les objectifs d'une opération par un questionnement scientifique précis ;

– réviser les connaissances après une opération ;

– comparer des opérations précédentes sur un même site ou dans un même territoire ;

– réaliser des études statistiques complètes de divers types.

2. Le travail d'appariement avec le modèle CIDOC-CRM

L'appariement est la mise en correspondance entre les entités d'un modèle de base de données à apparier et une classe du CIDOC-CRM considéré comme le modèle cible et les propriétés associées aux classes.

L'appariement des termes et des concepts est une étape essentielle pour vérifier la compatibilité des

1 christophe.tuffery@inrap.fr

2 emeline.legoff@inrap.fr

3 www.cidoc-crm.org et www.cidoc-crm.org/collaborations

modèles de systèmes de description d'entités d'un domaine culturel avec le modèle cible du CIDOC-CRM et de ses extensions. Les principes à suivre pour effectuer ce travail sont décrits par Kondylakis H. et al. (2006).

En 2015, dans le cadre du programme de formation Trans National Access du projet européen ARIADNE[4], nous avons pu réaliser un premier test d'appariement entre les classes et propriétés de l'extension CIDOC-CRMarchaeo et les tables de plusieurs systèmes numériques d'acquisition de données archéologiques de terrain[5]. Une présentation des résultats de ces tests a eu lieu lors du CAA2016 (Tufféry et al., 2016). D'autres tests d'appariement ont été entrepris au cours de l'année 2017.

Pour effectuer ce travail d'appariement, nous avons utilisé l'application 3M (Memory Mapping Manager), développée par l'Institute of Computer Science (ICS)[6]. Cet outil permet d'apparier des données archéologiques avec les classes et propriétés du modèle CIDOC-CRM et des extensions CRMsci, CRMarchaeo, CRMba spécifiques au domaine de l'archéologie, et d'autres extensions comme CRMgeo pour les données spatiales, CRMdig pour les métadonnées de représentations 2D, 3D ou animées. Concrètement, l'appariement avec 3M exploite un fichier .xml provenant des tables de données des systèmes d'enregistrement de terrain étudiés. Le passage des fichiers sources à leurs versions au format .xml s'appuie sur une exportation des tables au format .csv puis au format .xml avec un éditeur XML. Une validation de la syntaxe .xml est effectuée avec un outil en ligne[7] (fig.1).

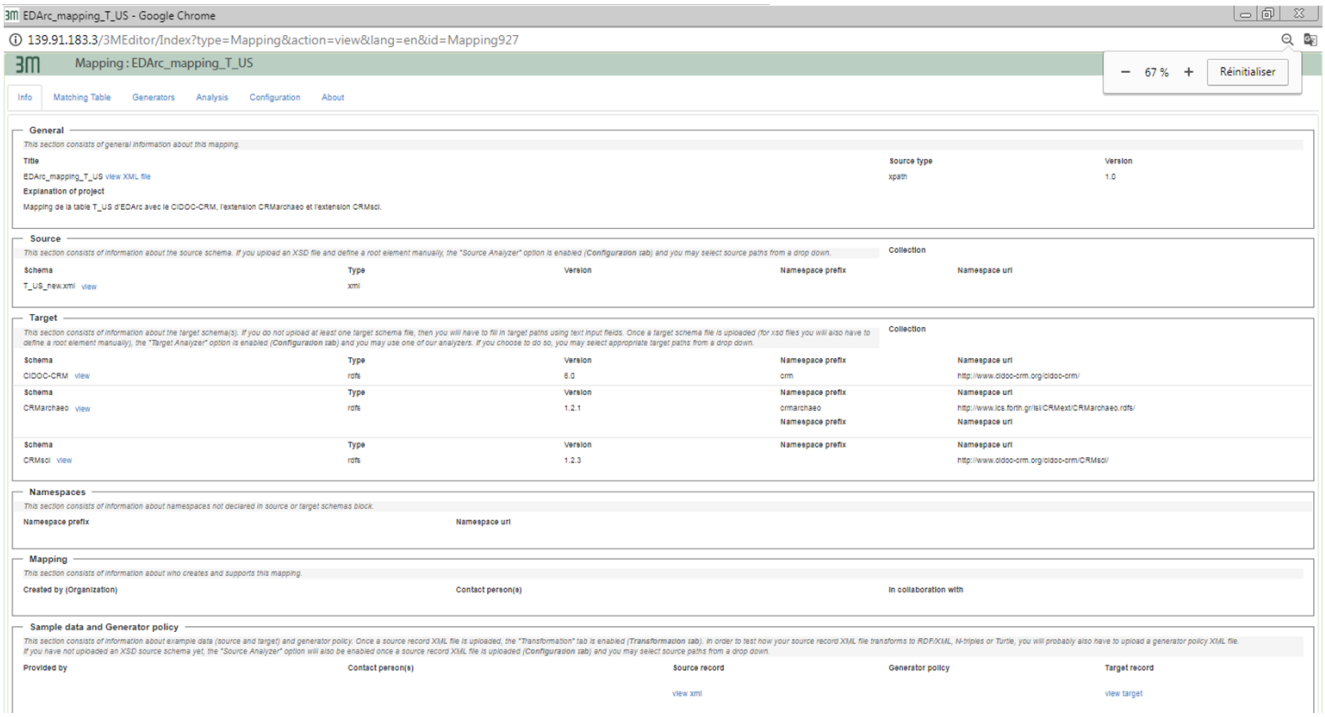

figure 1 : Interface de l'application 3M pour charger les fichiers .xml sources et cibles

Une fois le fichier .xml chargé dans 3M, il est possible d'apparier les champs des tables avec les classes et les propriétés adaptées du CIDOC-CRM et des extensions utilisées (fig.2).

4 http://www.ariadne-infrastructure.eu/

5 Ce travail a été réalisé avec les systèmes d'enregistrement de terrain ArcheoDB, utilisant MS Access, et CADoc, utilisant FileMaker Pro, respectivement développés par N. Holzem et T. Guillemard, tous les deux techniciens de fouille de l'Inrap que nous tenons à remercier

6 http://139.91.183.3/3M/

7 http://www.xmlvalidation.com

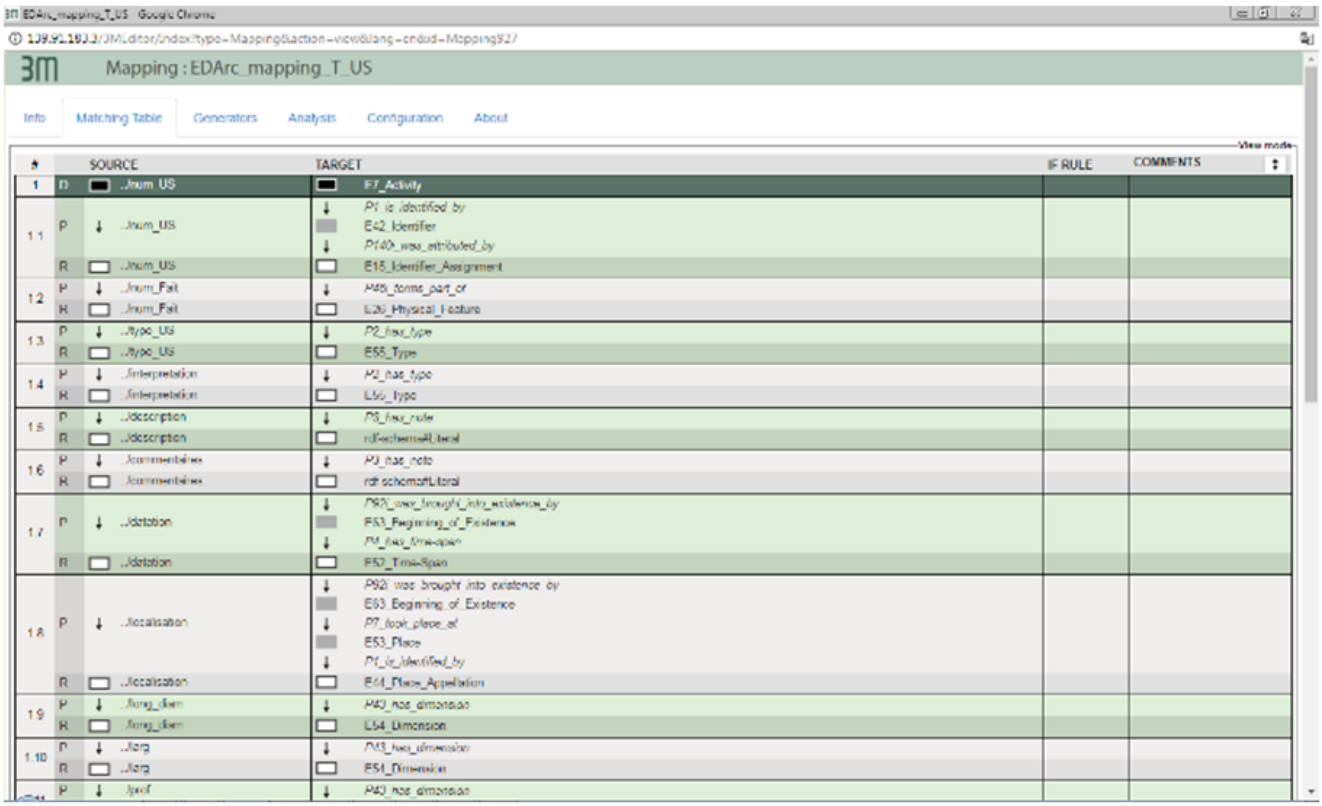

figure 2 : Interface pour réaliser l'appariement

3. Premiers résultats

L'une des premières conclusions des tests est qu'il n'existe pas une description unique possible des entités archéologiques et de leurs relations selon le modèle CIDOC-CRM. Cette diversité tient aussi bien aux choix faits pour désigner les entités archéologiques et les relations entre entités, qu'à la diversité des types d'opérations archéologiques.

Une opération archéologique doit être déclarée en utilisant la classe E7_Activity et une classe de localisation géographique E53_Place. La fouille archéologique elle-même doit être déclarée à l'aide des classes A9_ArchaeologicalExcavation, A10_ArchaeologicalEcvavationArea et A11_ExcavationAreaDefinition. La classe utilisée pour décrire les unités archéologiques stratifiées peut être la classe A8_StratigraphicUnit (fig.3).

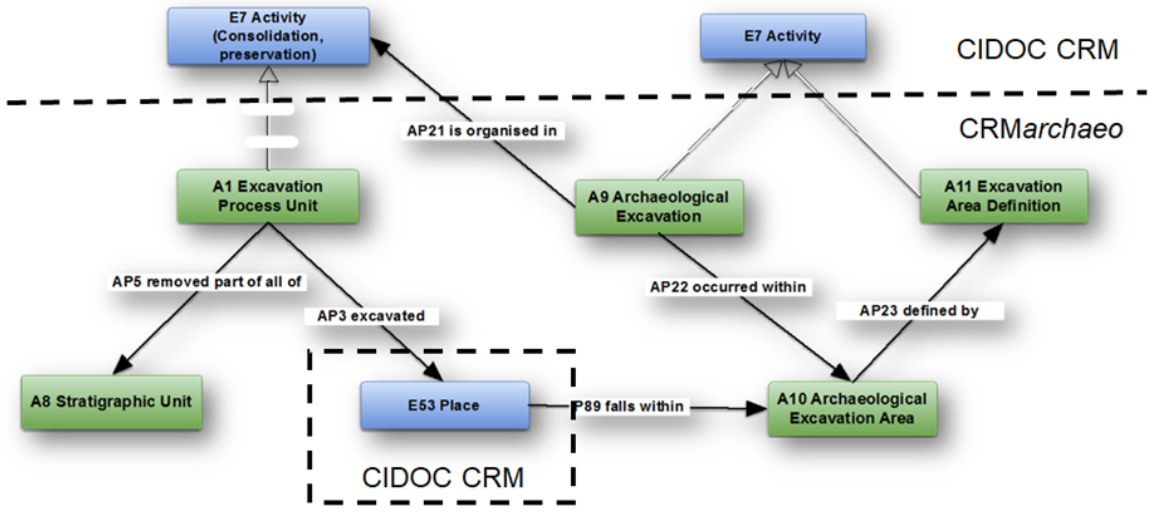

figure 3 : Classes principales d'une opération archéologique

Mais, on peut aussi faire un usage combiné des classes A3 (Stratigraphic Interface) et A2 (Stratigraphic Volume Unit), si on souhaite détailler les entités composant une unité stratigraphique. Un fait archéologique peut être décrit de la même façon que des unités stratigraphiques ou en décomposant les unités qui le constituent. Mais un fait peut aussi être décrit en utilisant les classes S20 (Physical Feature) et S10 (Material Substance) provenant du modèle de l'extension CRMsci (fig.4).

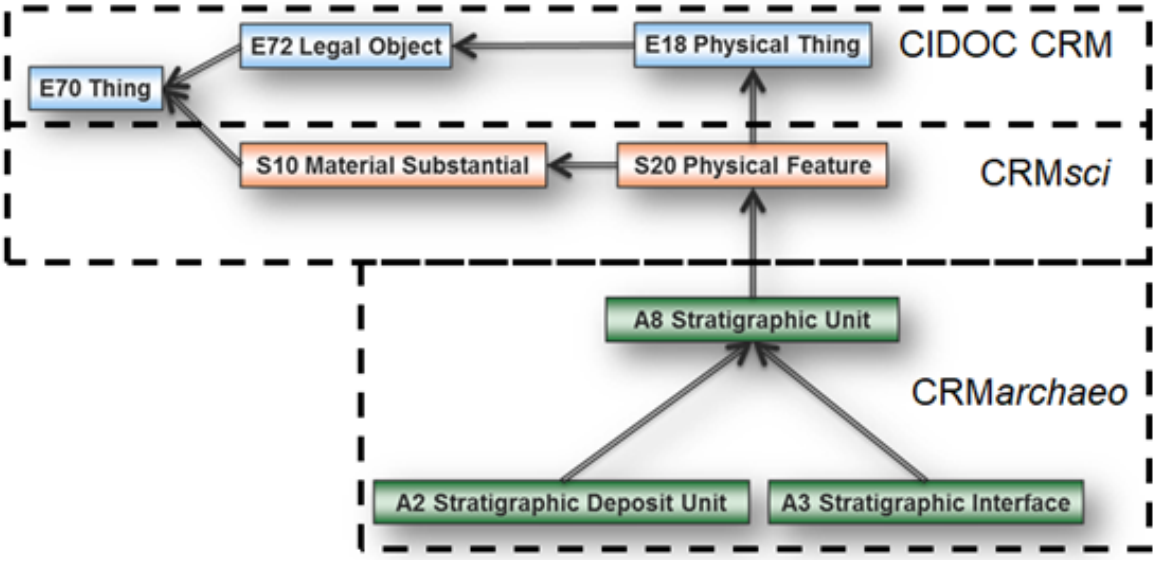

figure 4 : Classes spécifiques pour décrire une stratigraphie

Par ailleurs, les notions temporelles de chronologie absolue des unités stratigraphiques peuvent être décrites dans le modèle CRMarchaeo en utilisant diverses classes (E50Date, E63Beginning Of Existence, E52Time-Span, E64End of Existence, E92SpaceTimeVolume). Les notions de chronologie absolue utilisent la classe E49(TimeAppelation). Les composantes spatiales ont été décrites en utilisant des classes propres à la localisation absolue (E53Place et E27Site) ou les propriétés des relations de stratigraphie comme AP11(hasPhysicalrelation) et AP13(has stratigraphicrelation (is stratigraphic relation of)) (fig.5).

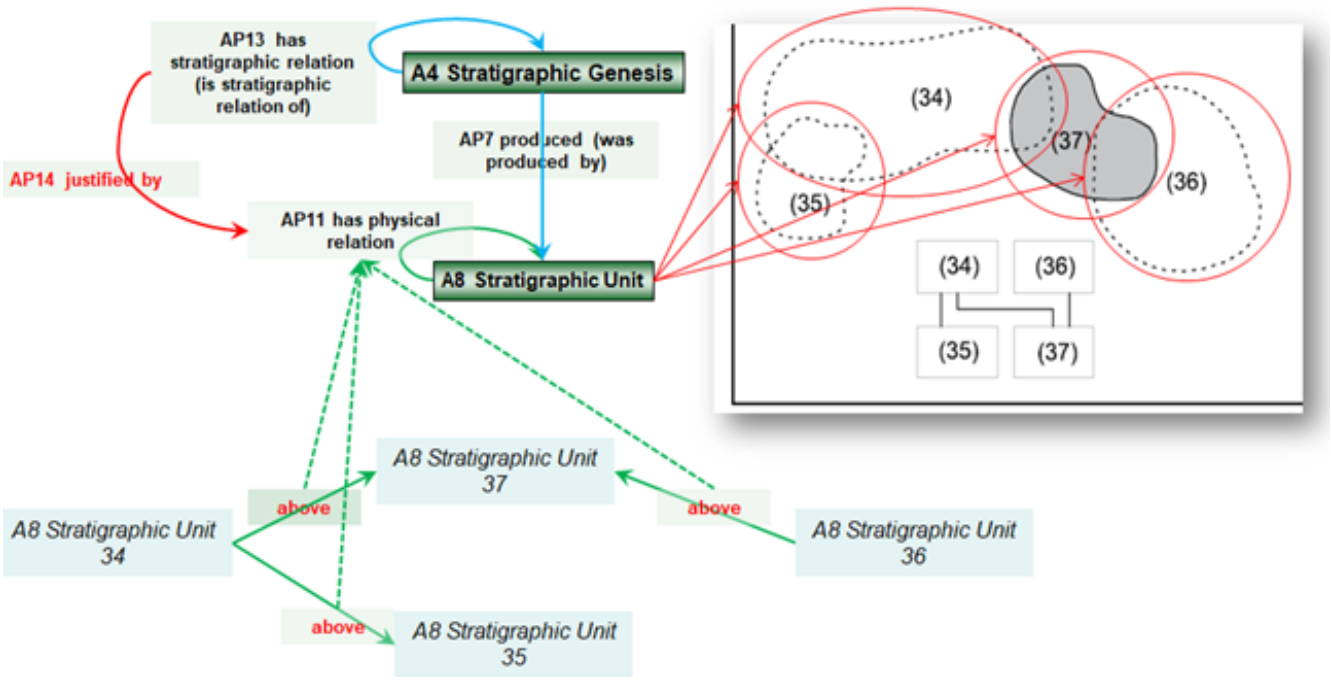

figure 5 : Classes spécifiques pour décrire les relations de chronologie relative entre unités stratigraphiques

4. Conclusions provisoire et suites envisagées

A ce jour, le modèle du CIDOC-CRM et les extensions utiles de ce domaine pour l'archéologie semblent adaptés aux types de données produites en archéologie préventive. Nous avons pu faire le constat qu'il n'existe pas une seule description possible des entités archéologiques et de leurs relations. Plusieurs appariements sont envisageables pour un même ensemble de données. Cette diversité tient aux choix faits pour désigner les entités archéologiques et les relations entre entités, à la diversité des types d'opérations et des contextes archéologiques, des problématiques scientifiques liées à l'enregistrement, des objectifs de publication, ou encore des points de vue des acteurs, etc. Cette marge de liberté dans le choix du mode de description des données provenant de divers systèmes d'enregistrement de terrain, comme cela a déjà été décrit par d'autres auteurs (E. Le Goff et al., 2015, M. Doerr et al. 2017). Il convient de souligner qu'il n'existe pas de procédure de validation des appariements ni d'autorité en charge de ce type de procédure. La validation reste donc à la libre appréciation des auteurs des appariements.

Les tests d'appariement seront poursuivis afin d'étendre la représentativité des résultats des premiers tests et de pouvoir confirmer la robustesse du modèle CIDOC-CRM pour un grand nombre de données archéologiques de l'Inrap sur diverses opérations (diagnostics et fouilles). En 2018, nous reviendrons sur les premiers tests d'appariement, nous ferons réaliser un même appariement par plusieurs personnes formées à cet exercice, nous réaliserons aussi l'appariement d'autres de bases de données de terrain utilisées à l'Inrap. Une fois les appariements validés, les données de terrain pourront être publiées dans le Web Sémantique Ouvert (Linked Open Data) sous forme de fichiers RDF[8] et en utilisant un Endpoint, exploitant le langage de requête et protocole SPARQL[9] propre aux données RDF. Mais cette prochaine étape nécessite que les données archéologiques de terrain soient publiées en licence ouverte, ce qui reste un sujet en cours de réflexion à l'Inrap et sur lequel aucune décision n'a été encore été prise.

discussion

AC : Anne Chaillou ; RD : Robert Demaille ; BD : Bruno Desachy ; JG : Julie Gravier ; CT : Christophe Tufféry)

BD : question provocatrice : à quoi ça sert ? On peut distinguer deux natures de l'information produite par les opérations archéologiques : la partie "référentiel" qu'on peut aussi appeler la "documentation primaire" : c'est à dire les inventaires (de contextes, de mobilier, etc.) avec leurs protocoles descriptifs définis, et la partie "discursive" qui contient les résultats et interprétations (argumentés et élaborés à partir de cette documentation "primaire"). Pour établir des synthèses au delà de chaque opération archéologique, c'est à dire pour fabriquer de la connaissance archéologique à la fois plus large et plus profonde (ce qui doit être le but, sauf à faire du formalisme un but en soi), on peut utiliser directement les conclusions du fouilleur - la partie discursive ; mais là, la formalisation exposée sert peu car elle porte sur la documentation primaire, pas sur le discours interprétatif (la meilleure formalisation de celui ci, dans la lignée des travaux de J.C. Gardin, est un enjeu intéressant pour de meilleures synthèses mais précisément ce n'est pas le champ du CIDOC CRM). On peut par ailleurs retourner à la documentation primaire pour des synthèses thématiques, en particulier statistiques : mais là il faut de toute façon composer ou recomposer des tableaux de données, c'est à dire revenir aux inventaires de base sous leur forme tabulaire d'origine (lignes de données et champs descriptifs). Alors quel est l'intérêt de traduire cette forme tabulaire du référentiel dans le formalisme et l'outillage logiciel spécifique de l'appariement avec CIDOC CRM ? Je crains même que les classes CIDOC CRM, à visée générale, ne "floutent" les données ; or si l'on veut faire un travail approfondi sur le référentiel, il faut revenir aux catégories définies par le fouilleur, quitte à les critiquer.

---

8 Resource Description Framework

9 SPARQL Protocol and RDF Query Language

CT : la démarche est un outil de clarification intellectuelle, qui permet d'explorer les concepts mobilisés dans l'enregistrement (unités et relations). En effet, l'appariement n'est pas verrouillé et n'oblige absolument pas à perdre des nuances spécifiques en faisant rentrer de force un système d'enregistrement dans des catégories prédéfinies rigides (ce qui est la fausse idée que l'on pourrait se faire de CIDOC CRM avant de vraiment rentrer dedans). Les extensions archéologiques de CIDOC CRM ne sont pas un cadre fixe ; l'opération d'appariement est au contraire une règle du jeu à la fois rigoureuse et souple, qui permet de mieux définir ses concepts propres d'enregistrement , en les confrontant avec les notions d'autres systèmes d'enregistrement, exprimées avec la même rigueur formelle, dans une démarche analytique et comparative. On peut ainsi voir si il n'y a pas déjà des utilisations de classes et de relations qui conviennent à ce que l'on veut exprimer, et sinon, il est possible d'en développer de nouvelles. Pour ma part, ce travail d'appariement, en particulier l'utilisation des triplets RDF (triplets sémantiques de type sujet/prédicat/objet, liés à la définition d'une ontologie et de ses classes) (note : voir notamment présentation de R. Demaille, compte rendu 2016-2017, séance du 18/02/2017, p.50) m'a été d'une grande utilité pour expliciter et clarifier les différents concepts d'enregistrement de terrain.

BD : Cette réponse me convainc presque de l'utilité intellectuelle du CIDOC CRM ! Ce que tu dis indique que le formalisme du modèle est un bon outil heuristique pour analyser et clarifier sa pensée et celle des autres. C'est incontestablement très important. D'autres formalisations de structuration de données peuvent aussi avoir ce rôle heuristique, notamment le modèle relationnel classique. Il est vrai que l'analyse de l'information en triplets sémantiques est sans doute plus aisée à s'approprier intellectuellement.

RD : le modèle relationnel est adapté à des domaines d'activités bien définis avec des concepts fixés, mais moins aux concepts non fixés ou encore en débat.

JG : je ne suis pas d'accord avec ce que dit Bruno sur le traitement statistique des référentiels : on ne doit pas nécessairement revenir au plus petit détail de catégorisation opérée par le fouilleur ; pour des synthèses statistiques larges il est intéressant au contraire de recourir à des catégories larges (pourvu qu'elles soient signifiantes bien sûr) ; dès lors, le travail d'appariement avec le modèle CIDOC CRM exposé par Christophe est une approche intéressante pour définir de telles catégories partagées. En outre, si un très grand nombre d'opérations sont appariées de cette manière, il est très probable que de légères différences sémantiques opérées par les archéologues sur les classes ne soient pas un problème car ces différences seraient statistiquement résiduelles. Toutefois (bien entendu), même avec un grand nombre d'opérations appariées, si les définitions des classes sont floues ou si les classes sont polysémiques, le traitement statistique des référentiels ne servirait à rien car on risque de ne faire ressortir que des différences entre des "écoles de pensée".

BD : C'est vrai, tu as raison sur l'utilité de traitements statistiques à un niveau global de catégorisation ; je nuance donc mon propos. Outre l'appariement de classes d'un modèle avec un autre, Christophe a évoqué dans sa présentation orale l'autre forme de "mapping" (mise en correspondance) qu'est l'alignement d'un thésaurus (ou lexique normalisé) avec un autre. A ce propos, j'ajoute que je reste partisan de l'idée selon laquelle plus les classes sont globales, moins le thésaurus mobilisé pour leur identification doit être détaillé ; notamment pour éviter cet écueil que tu mentionnes, de variations de termes choisis pour caractériser les occurrences de classes risquant de refléter des différences dialectales entre tribus d'archéologues, autant ou plus que des phénomènes archéologiquement signifiants.

AC (commentaire complémentaire post-séance) : thesaurus, hum ! mais bon c'est pas grave. ll serait intéressant de pouvoir faire un "mapping" CIDOC CRM des données d'échange (ou classeur Chaillou pour l'Inrap). On aura un système d'échange de données qui serait réellement normé et dont les champs seraient fixés par des triplets. Avantage : éviter le risque d'un possible "mapping" différent d'une même structure en fonction de la personne qui fait ce "mapping" (cf. le test de "mapping" d'une même structure par différentes personnes que Christophe souhaite mener en interne à l'Inrap). Le "mapping" des données d'échange devra être fait sur la nouvelle version après intégration des nouveautés législative de 2016 (travail en cours). A suivre ...

CT (commentaire complémentaire post séance) : A mon sens, il y a un intérêt à faire procéder à l'appariement avec le CIDOC de mêmes tables par plusieurs personnes, afin d'évaluer les réelles marges de manœuvre interprétatives que laisse le CIDOC. Dans l'usage de certaines classes du CIDOC et des triplets RDF associés, il est possible de laisser la place à des classes de type commentaires qui permettent de respecter des champs descriptifs d'interprétation des unités de terrain enregistrées. L'appariement avec le modèle cible qu'est le CIDOC permet de revenir sur le modèle source qu'est le système d'enregistrement et, le cas échéant, de faire évoluer sa structure et/ou ses champs descriptifs Il y a donc un effet retour ('feed back") intéressant. Cette démarche d'appariement a donc une dimension systémique. Le CIDOC n'est pas un "monde facile" à s'approprier pour la communauté des archéologues, peu habituée à ce type de formalisme et à l'usage de normes. Il faut donc développer à son attention des actions de formation/information sur le sujet (cf. Journées du réseau MASA sur l'interopérabilité qui ont eu lieu à Tours du 20 au 22 novembre 2017 : http://masa.hypotheses.org/430). Le projet d'Anne de faire l'appariement des classeurs d'échange parait indispensable. Cela peut permettre d'évaluer leur "robustesse" et aider à leur diffusion et leur adoption pour les systèmes d'enregistrement de terrain qui seraient appariés avec le CIDOC. Petite précision sémantique : le terme d'appariement (traduction de "mapping) désigne la mise en correspondance entre classes du CIDOC et champs des tables de données descriptives de systèmes d'enregistrement; le terme d'alignement doit être réservé à l'usage de thésauri et vocabulaires contrôlés (ex. PACTOLS). Ce genre de travail d'alignement a été réalisé notamment dans le cadre du programme européen ARIADNE (http://www.ariadne-infrastructure.eu/), en particulier pour le portail de ressources documentaires (http://portal.ariadne-infrastructure.eu/)

références citées

Doerr M., Felicetti A., Hermon S., Hiebel G., Kritsotaki A., Masur A., May K., Ronzino P., Schmidle W., Theodoridou M., Tsiafaki, D. and others. (2017). Definition of the CRMarchaeo An Extension of CIDOC CRM to support the archaeological excavation process. Proposal for approval by CIDOC CRM-SIG. Version 1.4. February 2016 :
http://www.ics.forth.gr/isl/CRMext/CRMarchaeo/docs/CRMarchaeo_v1.4.pdf

Kondylakis H., Doerr M., Plexousakis D. (2006) Mapping Language for Information Integration. Technical Report 385, ICS-FORTH, December 2006 :

http://www.cidoc-crm.org/sites/default/files/Mapping_TR385_December06.pdf

Le Goff E., Marlet O., Rodier X., Curet S., Husi P. (2015). Interoperability of the ArSol (Archives du Sol) database based on the CIDOC-CRM ontology. Actes du colloque CAA 2014, Paris, pp. 179-186

Tufféry C., Felicetti A., Jard P., Holzem N., Guillemard T. (2016). An essay of mapping archaeological land-record systems used by Inrap with CIDOC-CRM and CIDOC-CRMarchaeo extension using 3M on-line tool. Actes du colloque CAA 2016, S11-03, Oslo

Christophe Tufféry, Emeline Le Goff, Julie Boudry, Federico Nurra. Recours au CIDOC-CRM pour évaluer l'interopérabilité de données archéologiques de terrain très variées : présentation des premiers résultats des tests effectués par l'Inrap. Actes de SAGEO, Spatial Analysis and GEOmatics 2017, Novembre 2017, Rouen, France. hal-01649750 https://hal.archives-ouvertes.fr/hal-01649750/document

Christophe Tufféry, Emeline Le Goff, Contribution de l'Inrap à l'usage du CIDOC-CRM pour les données archéologiques d'enregistrement de terrain. Participation à la table-ronde des Journées MASA Interopérabilités, 20-22 novembre 2017, Tours : http://masa.hypotheses.org/430