



HAL
open science

Proposition méthodologique de gestion de la complexité et de l'hétérogénéité par le recours au web sémantique dans une perspective d'étude patrimoniale

Loic Jeanson, Florent Laroche, Jean-Louis Kerouanton

► To cite this version:

Loic Jeanson, Florent Laroche, Jean-Louis Kerouanton. Proposition méthodologique de gestion de la complexité et de l'hétérogénéité par le recours au web sémantique dans une perspective d'étude patrimoniale. 16ème colloque national AIP Primeca – s.mart, Les Karellis, Apr 2019, Les Karellis, France. hal-02498678

HAL Id: hal-02498678

<https://hal.science/hal-02498678>

Submitted on 4 Mar 2020

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.

Proposition méthodologique de gestion de la complexité et de l'hétérogénéité par le recours au web sémantique dans une perspective d'étude patrimoniale

Sortir du fonctionnement documentaire au risque de l'hétérogénéité.

Loïc Jeanson ^{1,2}

¹ CFV, ² LS2N

loic.jeanson@univ-nantes.fr

Florent Laroche ²

² LS2N | Centrale Nantes
Bâtiment S – 1, rue de la Noë
44321 Nantes – France
florent.laroche@ec-nantes.fr

Jean-Louis Kerouanton ¹

¹ Centre F. Viète (CFV) | U. Nantes
2 rue de la Houssinière - BP 92208
44322 NANTES – France
jean-louis.kerouanton@univ-nantes.fr

Résumé— Les données manipulées dans le cadre des modélisations ou plus généralement de l'étude historique ou patrimoniale d'objets de patrimoine culturel sont remarquables par leur grande hétérogénéité. Les ingénieurs et historiens/experts du patrimoine en font la vive expérience dans le cadre des travaux interdisciplinaires menés dans le cadre du projet ReSeed. L'objectif est d'atteindre un bon niveau d'intégration de ces données pour la production de connaissances nouvelles. La difficulté d'intégration a plusieurs causes. Le fonctionnement par fichier, le manque de connexion entre outils spécialisés, ce qui se résume dans la difficulté inhérente à gérer une grande hétérogénéité et variabilité de la nature des informations en jeu. Ce problème d'hétérogénéité pour l'intégration tel qu'il est rencontré dans le cadre de ReSeed est exposé avec quelque détail explicitant le cadre dans lequel nous nous plaçons. Il est suivi de notre proposition de réponse méthodologique, précisée dans l'expression d'éléments d'implémentation spécifiques. Enfin, parce qu'il nous semble qu'au bout du compte, notre réponse au problème nous aura surtout permis de mieux préciser la nature de la difficulté, une dernière partie présente les limites de la réponse et les pistes de réponses à venir.

Mots-clés— web sémantique, intégration, hétérogénéité, interdisciplinarité,

I. INTRODUCTION

Le projet ANR ReSeed réunit autour d'une même table des acteurs institutionnels et industriels dans l'optique de proposer des solutions nouvelles pour outiller les processus de compréhension des objets de patrimoine culturel. De plus amples informations sont disponibles sur le site internet du projet : <http://www.reseed.fr>. Chercheurs en génie mécanique ou en génie industriel de l'UTT, de l'UTC, de l'École Centrale de Nantes et de DeltaCAD, et historiens et conservateurs de patrimoine de l'Université de Nantes, de MCC Heritage, de la mission Inventaire général de la direction des patrimoines du

Ministère de la Culture français et du Musée des Arts et Métiers dialoguent pour définir des besoins et des approches pour des outils facilitant 1) la structuration d'informations et de données descriptives de ces objets de patrimoine culturel et de leurs contextes, 2) les activités de rétro-conception à partir de l'existant.

C'est dans ce cadre que ces travaux de recherches se déroulent, avec une attention particulière donnée à trois objets de patrimoine culturel :

- L'observatoire du Pic du Midi, plateforme scientifique de haute montagne dans les Pyrénées françaises
- Une collection de lunettes méridiennes, installées dans différents observatoires astronomiques en France. Certaines seulement sont encore en fonctionnement, l'une d'entre elle est démontée et rangée dans des réserves.
- Une maquette de la maison coloniale métallique produite par la maison Moreau frères.

Ces trois objets d'étude, apportés par des partenaires différents voient se poser sur eux des attentes et des méthodologies différentes, des usages et des concepts différents. L'intention est de proposer un outil suffisamment modulaire pour qu'il n'en résulte pas en la production de trois outils correspondant aux trois usages. Cependant, l'ambition n'est pas non plus que l'outil résolve tous les problèmes, qu'il puisse "tout faire".

Ainsi, dans ce cadre 'et en même temps', nous présenterons dans ce papier en trois parties notre problématique délimitée et en détails, ensuite comment nous y avons répondu et enfin nous jetterons notre regard sur ce qu'il reste des limites premières, sur les nouvelles contraintes que nous avons apportées et sur notre approche épistémologique de l'outil.

II. DES OBJETS DEFINIS PAR LEUR ETUDE

L'intention de modélisation (au sens de Vallespir) [1] derrière l'étude de ces objets est toujours sensiblement la même : expliquer la valeur patrimoniale de l'objet ou des objets. Cela passe, d'une part, par la construction d'une compréhension fine et nuancée de l'objet et d'autre part par une restitution de cette même compréhension dans l'espoir de sa transmission.

L'observatoire du Pic du midi, qui est le cas d'étude se lequel nous nous sommes penchés en premier, cherche à monter un dossier de candidature solide, afin de revendiquer la valeur du site auprès de l'UNESCO au titre du patrimoine mondial. ICOMOS, l'association s'occupant d'instruire les dossiers relevant du patrimoine culturel pour l'UNESCO a explicité une méthodologie plus ou moins précisée qui permet de couvrir une grande diversité de réalités avec une méthodologie commune[2]. Avec cette méthodologie comme guide, nous cherchons à modéliser, à construire une représentation de l'authenticité et de l'intégrité du site, et à permettre une analyse comparative qui mettra en lumière les qualités du site au regard des qualités des autres sites. Le site a beaucoup évolué depuis la fin du 19^e siècle où les premiers bâtiments ont été construits. Des activités scientifiques variées y ont eu lieu : astronomie, botanique, aérologie, météorologie, etc. La complexité de l'objet réside dans la grande diversité des variables qu'il faut considérer : lieux, bâtiments, objets, acteurs, activités scientifiques, activités touristiques, résultats de recherche, ... Certaines activités étaient partagées entre plusieurs observatoires, certains acteurs partageaient leurs temps entre plusieurs activités, plusieurs institutions, certains instruments ont servi à plusieurs activités, etc. Au-delà de la quantité et de la diversité, l'enjeu réside dans la granularité qu'il est nécessaire de produire.

La collection de lunettes méridiennes change l'échelle. Les lunettes, ne représentent, en effet, que des volumes de l'ordre de "quelques mètres cubes" quand l'observatoire du Pic du midi représente plusieurs milliers de mètres cubes. Leur approche n'en est pour autant pas simplifiée, car elles apportent un niveau de complexité nouveau. Comme tous les objets furent construits avec une intention de standardisation[3], l'intention est d'étudier la variabilité 1) par rapport au modèle initial, 2) par la 'vie' de l'objet, au travers de son usage. La mise en série permet de distinguer ces variations par rapport à des modèles de référence inexistantes (les constructeurs travaillaient sans plan), que l'on s'attachera pourtant à dégager. Une étude mécanique fine est également dans le cahier des charges.

La maquette de la maison coloniale Moreau à commencer au courant de l'année 2019 sera l'occasion d'une étude des formes et des typologies de la structure et des assemblages. L'un des intérêts supplémentaires est le lien entre la maquette et les objets physiques : comme une maquette numérique, la maquette physique permet de mieux comprendre la maison produite et vendue par la compagnie Moreau frères.

Le lien entre toutes ces approches réside dans l'intention de montrer la valeur patrimoniale. Le résultat final (i.e. une/des façons d'appréhender et d'expliquer cette valeur patrimoniale) n'est jamais certain avant d'avoir commencé l'étude des objets et

de leur contexte. Ces trois objets montrent également une bonne dose d'hétérogénéité qu'il faut appréhender. Elle passe en premier lieu par leur échelle. Du site étendu sur plusieurs centaines de mètres carrés, à la maquette de musée de quelques centimètres cubes, la marche est grande. Au sein de cette dimension d'échelle la granularité rajoute de la complexité, en liant les échelles, en redécoupant les entités et les réassemblant.

Mais l'hétérogénéité réside également dans l'approche méthodologique, explicite, mais incluant des zones de plus ou moins libre interprétation, se définissant au fil de l'eau de l'étude, peu basée sur des éléments d'analyse factuelle en tous points communs. Ces éléments factuels, constitutifs de la modélisation ou de l'étude, apportent leur lot d'hétérogénéité : nom, géométrie, matériau, texture, usage, propriétaire, lieu d'implantation, date de découverte, etc. Les divergences fines entre différentes informations (des dates, par exemple : date de création, date d'ouverture, date de découverte, ...) causent une multiplicité des paramètres à considérer, comme autant de types d'information pour décrire et comprendre l'objet. En revers, les modélisateurs aussi, s'adaptant à leurs conceptions ontologiques des objets apportent des divergences fines. Pour reprendre l'exemple des dates : comment définir la date de création d'un tableau ? Certains considéreront la date d'inauguration de l'exposition révélant le tableau au public pour la première fois, d'autres la journée supposée de fin de création de l'œuvre (introduisant alors une variable d'incertitude), d'autres encore la datation de la lettre dans laquelle le peintre fait référence à ce tableau comme achevé. Toutes ces dates donnent des indices plus ou moins certains et de points de vue ontologiques différents sous un même label standardisé de 'date de création'. Quelle valeur porte alors le recoupement ou la comparaison de multiples 'dates de création' (porteuses de différences ontologiques dont on perd la trace) ?

Et au sein même d'un type, persiste la nécessité de distinguer encore d'autres notions. Comme par exemple, celles de modèle et d'implémentation : la forme d'un objet diffère de la géométrie portée par le plan de construction. On peut aussi envisager l'influence du temps et de l'usage sur la géométrie [4] : la forme d'une pièce de fonderie, vieille de 150 ans, montée sur une machine, diffère de la forme de son moule ou d'une pièce sœur de rechange, stockée pendant le même temps.

Les types d'hétérogénéité dans les informations à considérer commencent à s'accumuler et notre liste est très loin d'être exhaustive ou de concerner tous les métiers, toutes les applications, tous les objets matériels de patrimoine culturel (nous n'ouvrons pas ici le champ du patrimoine immatériel).

En peu de mots, notre problème d'hétérogénéité se pose par un défaut d'outil capable de gérer correctement tous ces types de données, toutes ces notions et ces échelles séparées et pourtant liées les unes aux autres. Des logiciels de CAO ou plus généralement de création de modèles 3D manipulent les formes, parfois les échelles, les textures, la contraintes cinématiques, etc. Les logiciels de gestion de bases de données manipulent correctement des données pour peu qu'on ait défini un schéma, un méta-modèle et que l'on contrôle l'hétérogénéité de nature des données (i.e. si un booléen est attendu, s'assurer que les données ne soient que des booléens, etc.). La spécialisation des outils

répond à des besoins spécifiques mais s'oppose à notre besoin d'intégration forte.

III. NOTRE PROPOSITION, OU COMMENT METTRE SUR UN MEME PLAN DES ELEMENTS PROFONDEMENT DIFFERENTS

Pour dépasser l'enclavement des outils, nous avons eu besoin de revenir à l'unité de base manipulée : les données, que nous interconnectons pour former des informations. Par la mise en série de ces informations, nous sommes en mesure de dégager des tendances, des schémas, des motifs : des connaissances. [5]

Ces données peuvent-être des coordonnées de points dans un espace doté d'une métrique, des dates, des noms de lieux, de personnages, etc. Cela ne reste que des données / informations / connaissances.

d'utiliser des outils de web sémantique qui permettent d'utiliser des vocabulaires ouverts et liés [6]. Ces vocabulaires conditionnent les relations entre données ou les données qu'il est possible de relier. Ils permettent de traiter dans un seul système des données très hétérogènes, de produire des 'exports' de ces données, d'intégrer des données depuis des bases extérieures. La méconnaissance du schéma de données d'une base distante n'empêche pas de la questionner (elle limite simplement notre capacité à la requête a priori). En résumé, les technologies du web sémantique nous offrent une solution flexible pour gérer des données, même très hétérogènes.

B. Un système réticulaire

Avec ce système de bases de données orientées graphes, ouvertes et liées, nous sommes en capacité de gérer des données, d'en extraire certaines pour les visualiser. Mais nous sommes assez loin de pouvoir les manipuler avec autant de puissance que

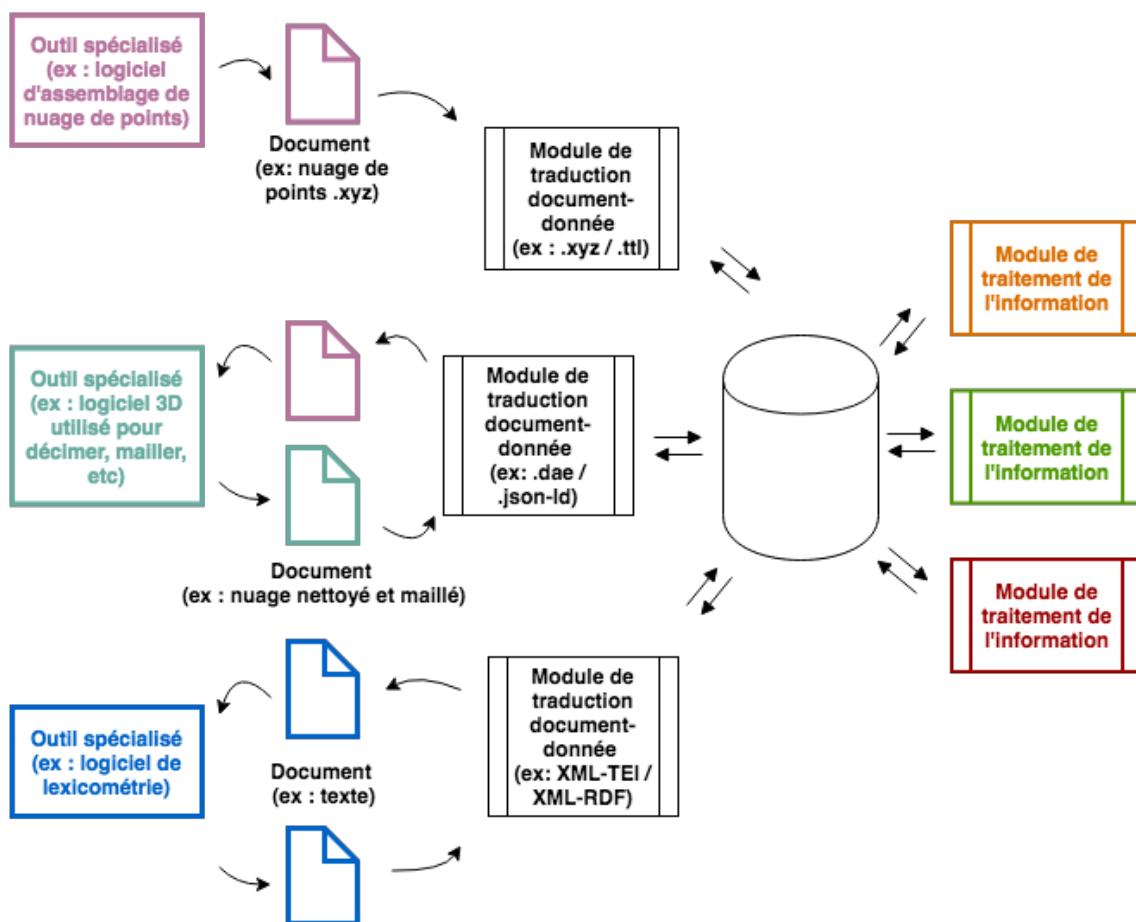


Figure 1 : Schéma de principe de l'interface outils spécialisés – base de donnée

A. Le web sémantique pour réduire les effets de la standardisation

Pour gérer ces données de façon uniforme et dépasser la logique de documents épars, nous faisons la proposition

de tous les outils spécialisés classiques. L'objet de notre proposition n'est cependant pas de les remplacer, simplement de pouvoir lier et analyser données, informations et connaissances que nous souhaitons manipuler. Notre proposition est celle d'un système réticulaire, au sens de Simondon[7], fait des outils

spécialisés rejoints par de cette couche 'liante', exposé dans la Figure 1.

Ce schéma de principe montre notre vision de la gestion des données dans le cadre du projet ReSeed : les données sont liées et interconnectées et peuvent être manipulées et analysées par différents modules de traitement de l'information. Autrement dit, une fois les données intégrées dans la base, certains traitements sont possibles. Cependant, ces données peuvent être importées ou exportées pour interaction avec d'autres outils spécialisés. Nous proposons, au fond, une sorte de marmite dans laquelle nous sommes en mesure de mélanger des ingrédients très différents, dans laquelle nous sommes en mesure de faire un certain nombre d'opérations. Mais depuis laquelle il est également possible de trier, hiérarchiser, extraire nos informations et d'en proposer une forme de traçabilité (qui reste cependant encore imparfaite).

C. Quelques développements propres à notre usage

Notre propre usage a poussé à produire des vocabulaires et des méthodes qui répondait à des besoins inassouvis dans ce que nous avons pu identifier. D'abord, la méthode historique, qui est en cela également celle des experts du patrimoine, nous a poussés à développer une systématique de la traçabilité. La scientificité de notre modèle dépend de sa résilience à la critique. Ainsi, donc dans un premier temps, l'accent sera mis sur les mécanismes de traçabilité qu'il nous aura fallu mettre en place. Dans un second temps nous verrons notre besoin de vocabulaires, pour la 3D d'abord et pour d'autres usages précis enfin.

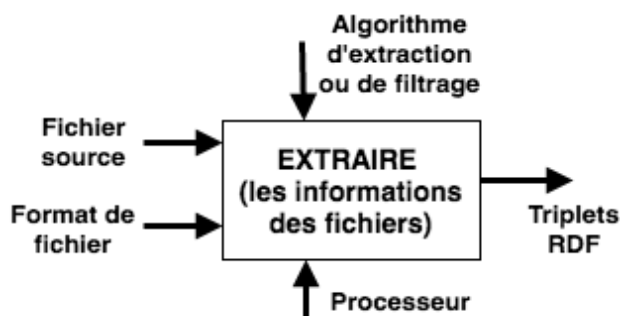


Figure 2 : Actigramme des opérations de traduction document - données

1) Processus de traçabilité

La traçabilité, telle que nous l'avons envisagée dans nos travaux couvre : les données et des informations, et les mécanismes de modification de l'information. L'idée est de produire un modèle aussi résistant à la critique que possible : la validité de l'information dans notre base est sujette à critique au même niveau que l'information telle qu'exposée dans la source et de même, les choix de modification de l'information sont tracés et explicités afin de permettre une critique des méthodes autant que des données produites.

a) Id pour les triplets

Les données sont stockées en triplets RDF. Nous avons, dans la pratique, détourné légèrement un outil du web sémantique : les *named graphs*. Utilisé pour délimiter un groupe de triplets en particulier, un *named graph* est, en quelque sorte, un sous-graphe du graphe général de notre base de données orientée graphes. Nous identifions chaque triplet par un *named graph*, dont le 'nom' est le *hash* du triplet. Ainsi chaque donnée et chaque information possède une URI : nous sommes en capacité de pister toute information dans notre base, depuis son processus de création et au travers de ses différentes altérations. Par le recours au *hash*, nous évitons de potentielles déduplications. Cette approche est inspirée des travaux remarquables du consortium Software Heritage[8].

b) Système de balisage

Tant que faire se peut, les données sont issues de sources. Très tôt, il nous aura fallu des outils pour baliser la position des données dans les sources afin d'affiner notre capacité à la traçabilité. Telle information n'est pas liée simplement à sa source, mais si elle est distinguable clairement, la position de l'information dans la source est elle aussi accessible. Ce balisage vaut pour tout type de source, que ce soit des textes, des images, des fichiers 3D, etc. Par exemple. Cette approche est inspirée des travaux et recommandations largement adoptés que fait le consortium IIF pour gérer les images en 2D[9].

c) Gestion de l'hypothèse et de la contradiction

Puisque l'objectif est de produire un modèle le plus complet possible, mais que le temps faisant son œuvre, l'information est toujours partielle : le propre des modèles historique et patrimonial est qu'il a recours à l'hypothèse. Il nous faut permettre d'intégrer des hypothèses et pire encore, d'intégrer des points de vue contradictoires. Les hypothèses ne sont pas fantaisies et peuvent être elles-mêmes issues de faisceaux plus ou moins denses d'indices et les points de vue contradictoires sont le terrain de discussion de travail, d'élaboration de nouvelles hypothèses. Le système de gestion de l'hypothèse est assez simple : la création d'un vocabulaire spécifiant la nature d'une information dans la base (issu de source ou produit par une hypothèse), de plus une hypothèse peut avoir pour source des informations ou des sources et la méthode de production d'une hypothèse est tracée.

2) Vocabulaires pour la 3D

Ce manque n'en est qu'un à moitié. Nous utilisons un vocabulaire existant (3D ontology) [10], qui est un alignement du schéma de données de X3D. Travaillant plus communément avec des fichiers Collada, notre usage n'est pas tout à fait satisfaisant. Il nous faut approfondir la chose, mais la représentation 3D d'objets reste une pratique qui nous semble assez peu en usage pour les utilisateurs des technologies du web sémantique (alors qu'elle nous semble précisément permettre le dépassement de la segmentation modèle 3D/base de données).

3) Vocabulaires manquants / adaptés à des usages précis.

Enfin, plus généralement, la pratique patrimoniale mène à constater qu'un très grand nombre de vocabulaires sont développés pour des applications spécifiques : combien de dérivés du CIDOC-CRM [11] ? Combien d'ontologie basée sur

des ontologies de haut niveau et ajustées à un métier, à une communauté ? La flexibilité des outils du web sémantique permet une diversité de vocabulaires, concomitants, se chevauchant, disjoints. Nous avons nous même proposé un vocabulaire permettant de décrire l'intégrité et l'authenticité d'après les définitions des textes fondateurs de l'UNESCO et recourrons à bon nombre de vocabulaires plus ou moins en développement, pour beaucoup n'étant pas des standards autres que locaux et n'ayant pas nécessairement beaucoup plus d'ambition.

IV. LIMITES ET PERSPECTIVES, OU L'HETEROGENEITE QUI SURVIT

Avec ce système, nous sommes en capacité de manipuler, de lier et d'intégrer toutes les données variables, variées et complexes avec lesquelles nous sommes amenés à travailler. La flexibilité répond à notre besoin d'affronter la complexité et l'hétérogénéité. Cependant, elle ne résout pas tout : l'hétérogénéité n'est que repoussée d'une étape. En effet, ce qui devient alors très hétérogène, ce sont nos vocabulaires et nos modèles de données.

Notons également que bon nombre d'ontologies ne sont pas définies logiquement (au sens de 'logique de premier ordre', par exemple) mais simplement sur des bases lexicales. Plusieurs initiatives sont menées afin de procéder à des travaux d'alignement, de 'mapping', d'ontologies entre elles.

Ces alignements permettent de retrouver un niveau de variabilité réduit dans les vocabulaires, voire dans les données utilisées. Cette réduction de la variabilité répond à un besoin d'efficacité à la requête, afin de ne pas passer à côté d'informations simplement structurées par une autre ontologie. Cette variabilité nous semble, cependant, une force et peut-être inhérente à la pratique de l'histoire et de l'étude patrimoniale : la valeur patrimoniale, le décisif tient dans les nuances, les petites différences qu'il serait dommageable de lisser. Nous nous trouvons dans une forme d'impasse, tiraillés entre notre besoin de variabilité, de faire survivre l'hétérogénéité malgré tout pour appréhender la nuance et une grande diversité de point de vue, de culture métiers et un besoin évident et pragmatique à l'efficacité de la requête. A quoi bon un système de base de données très complet et en mesure d'intégrer des données très hétérogènes s'il est inutile parce qu'il est impossible de le questionner correctement ? A cette question nous n'apportons aujourd'hui pas encore de réponse claire, mais nous travaillons sur le problème, en partant de ce constat et des travaux de recherche de Matthieu Quantin : Haruspex[12].

V. CONCLUSION

Ce papier présente notre approche pour l'intégration des données, malgré l'hétérogénéité. Après avoir décrit notre problème à partir de nos cas d'application, nous précisons rapidement notre réponse à cette difficulté et quelques éléments de pratique ajustée à la pratique historique et patrimoniale.

Enfin, nous mettons en avant un problème sous-jacent qu'il nous semble encore plus utile de résoudre pour permettre à notre solution d'être effectivement pertinente et utilisable.

VI. REFERENCES

- [1] B. Vallespir, D. Chen, et Y. Ducq, « ENTERPRISE MODELLING FOR INTEROPERABILITY », *IFAC Proc. Vol.*, vol. 38, n° 1, p. 70-75, 2005.
- [2] C. Cameron et M. Rössler, *La Convention du patrimoine mondial: La vision des pionniers*. Presses de l'Université de Montréal, 2017.
- [3] F. L. Guet-Tully et J. Davoigneau, « L'inventaire et le patrimoine de l'astronomie : l'exemple des cercles méridiens et de leurs abris », *Situ*, n° 6, sept. 2005.
- [4] J. Gleick et F. Bouillot, *L'information: l'histoire, la théorie, le déluge*. Paris: Cassini, 2015.
- [5] C. Zins, « Conceptual approaches for defining data, information, and knowledge », *J. Am. Soc. Inf. Sci. Technol.*, vol. 58, n° 4, p. 479-493, févr. 2007.
- [6] F. L. Gandon, C. Faron-Zucker, et O. Corby, *Le Web sémantique: comment lier les données et les schémas sur le Web?* Paris: Dunod, 2012.
- [7] G. Simondon et N. Simondon, *Du mode d'existence des objets techniques*, Nouv. éd. rev. et corr. Paris: Aubier, 2012.
- [8] R. Di Cosmo et S. Zacchioli, « Software Heritage Archive ».
- [9] S. Snyderman, R. Sanderson, et T. Cramer, « The International Image Interoperability Framework (IIIF): a community & technology approach for web-based images », in *Archiving Conference*, 2015, vol. 2015, p. 16–21.
- [10] Web3D Consortium, « 3D Ontology ».
- [11] CIDOC, « CIDOC CRM ».
- [12] M. Quantin, B. Hervy, F. Laroche, et J.-L. Kerouanton, « Haruspex, Outil de Gestion de Connaissances non Structurées », *Archéologies Numér.*, 2017.