Big Data and Archaeology

Proceedings of the XVIII UISPP World Congress
(4-9 June 2018, Paris, France)
Volume 15

Session III-1

edited by

François Djindjian and Paola Moscati

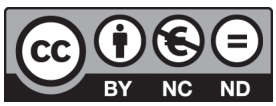
Archaeopress Archaeology

Archaeopress Publishing Ltd
Summertown Pavilion
18-24 Middle Way
Summertown
Oxford OX2 7LG

www.archaeopress.com

ISBN 978-1-78969-721-6
ISBN 978-1-78969-722-3 (e-Pdf)

This book is available direct from Archaeopress or from our website www.archaeopress.com



UISPP PROCEEDINGS SERIES VOLUME 15 – Big Data and Archaeology

UISPP XVIII World Congress 2018

(4-9 Juin 2018, Paris)

Session III-1

VOLUME EDITORS: François Djindjian and Paola Moscati

SERIES EDITOR: The Board of UISPP

SERIES PROPERTY: UISPP – International Union of Prehistoric and Protohistoric Sciences

KEY-WORDS IN THIS VOLUME: archaeology, archaeological theory, archaeological computing, Big Data

UISPP PROCEEDINGS SERIES is a printed on demand and an open access publication,
edited by UISPP through Archaeopress

Foreword to the XVIII UISPP Congress Proceedings

UISPP has a long history, starting 1865 with the International Congress of Prehistoric Anthropology and Archaeology (CIAAP), until 1931, date of the foundation in Bern of UISPP. In 1955, UISPP became member of the International Council of Philosophy and Human Sciences, associate of UNESCO.

UISPP is structured on more than thirty scientific commissions which a very representative network of worldwide specialist of prehistory and Protohistory, and covering all the specialties of archaeology: historiography, archaeological methods and theory; material Culture by period (palaeolithic, Neolithic, bronze age, Iron age) and by continents (Europe, Asia, Africa, Pacific, America), palaeoenvironment and palaeoclimatology; archaeology in specific environments (mountains, desert, steppes, tropical area), archeometry; Art and culture; Technology and economy; biological anthropology; funerary archaeology; Archaeology and societies.

The UISPP XVIII° world congress of 2018, in Paris, France with the strong support of all the French institutions related to archaeology, involved 122 sessions, over 1800 papers from scientist of almost 60 countries from all continents.

The proceedings, edited in this series but also as special issues of specialized scientific journals, will remain as the most important outcome of the congress.

L'UISPP a une longue histoire, à partir de 1865, avec le Congrès International d'Anthropologie et d'Archéologie Préhistorique (C.I.A.A.P.), jusqu'en 1931, date de la Fondation à Berne de l'UISPP. En 1955, l'UISPP est devenu membre du Conseil International de philosophie et de Sciences humaines, associée à l'UNESCO. L'UISPP repose sur plus de trente commissions scientifiques qui représentent un réseau représentatif des spécialistes mondiaux de la préhistoire et de la protohistoire, couvrant toutes les spécialités de l'archéologie : historiographie, théorie et méthodes de l'archéologie ; Culture matérielle par période (Paléolithique, néolithique, âge du bronze, âge du fer) et par continents (Europe, Asie, Afrique, Pacifique, Amérique), paléoenvironnement et paléoclimatologie ; Archéologie dans des environnements spécifiques (montagne, désert, steppes, zone tropicale), archéométrie ; Art et culture ; Technologie et économie ; anthropologie biologique ; archéologie funéraire ; archéologie et sociétés.

Le XVIII° Congrès mondial de l'UISPP en 2018, à Paris en France, avec le soutien de toutes les institutions françaises liées à l'archéologie, comportait 122 sessions, plus de 1800 communications de scientifiques venus de près de 60 pays et de tous les continents.

Les actes du congrès, édités par l'UISPP comme dans des numéros spéciaux de revues scientifiques spécialisées, restera comme le résultat le plus important du Congrès.

Marta Arzarello

Secretary-General /
Secrétaire générale UISPP

Contents

List of Figures

P. Moscati: **How Big Is Big Data?**

S. Grégoire *et al.*: **Innovative multidisciplinary method using Machine Learning to define human behaviors and environments during the Caune de l'Arago (Tautavel, France) Middle Pleistocene occupations**

F. Peudon *et al.*: **Cagny-l'Épinette (Somme Valley, France), Thirty Years of Mixed Data: Potential and Limits**

R. Roure *et al.*: **Towards an Archaeological Information System: the evolution of *Syslat*, an archaeological data management software**

F. Djindjian: **L'archéologie néoprocessuelle**

P. Gaussein: **Transcending 'Technocomplexes'. When French Empiricism calls for Hypothetico-deductive Method**

Introduction au volume

Le présent volume contient les articles de communications données dans deux sessions du XVIII Congrès mondial de l'UISPP, qui s'est tenu à Paris en juin 2018, organisées par la Commission « Méthodes et théorie de l'archéologie » :

- Session III-1 (CA) : Big data, databases and archaeology
- Session III-1 (T) : New advances in theoretical archaeology

Le volume contient six articles de la première session et deux articles de la deuxième session.

Après une introduction au thème des Big Data en archéologie donnée par François Djindjian, deux articles par Paola Moscati et Olivier Buchsenschutz retracent le contexte historiographique du thème et ses développements actuels. En suivant un laps du temps qui s'écoule des années 1950 à nos jours, le sujet est abordé d'un point de vue terminologique, conceptuel et technique, sous l'angle de la collection et du traitement de l'information et de l'interprétation et de la transmission de la connaissance.

Les articles de Sophie Grégoire (*et al.*) sur le site de la Caune de l'Aragon et de Floriane Peudon (*et al.*) sur le site acheuléen de Cagny l'Épinette, donnent des exemples complets d'application des méthodes informatiques (bases de données, systèmes d'information géographique) et d'algorithmes d'intelligence artificielle par apprentissage. Il en résulte une approche multidisciplinaire et intégrative à l'enregistrement, la gestion et l'exploitation des archives de fouilles et à la classification de la documentation, manuscrite et numérique, produite pendant une longue période durant laquelle l'archéologie a vu sa progressive informatisation.

L'article de Réjane Roure (*et al.*) concerne les plus récents développements du logiciel Syslat qui est un progiciel intégré d'acquisition et de gestion des données archéologiques, initialement développé dans les années 1980 pour le site de Lattes en Languedoc. Créé par Michel Py (CNRS), Syslat est aujourd'hui une suite d'outils complets et gratuits, destinée à l'enregistrement, l'exploitation et l'analyse documentaire des données de fouille et complétée par une application pour mobiles, une application web destinée à la consultation et un outil multiplateforme pour faciliter la saisie des données sur le terrain.

La deuxième session contient deux articles. Le premier, de François Djindjian, présente les caractéristiques générales d'une nouvelle plateforme théorique, l'archéologie néoprocessuelle, qui est une approche centrée sur le rôle des processus dans les principaux domaines de l'archéologie. Le second article, de Pascaline Gaussein, est une réflexion sur le concept de cultures archéologiques et sur les réalités humaines et sociales qui sous-tendent les cartes de répartition des traits culturels paléolithiques, en combinant les approches processuelle, contextuelle et empirique.

François Djindjian et Paola Moscati

Mégadonnées et archéologie : une introduction

François Djindjian
Université de Paris 1 Panthéon Sorbonne et CNRS UMR 7041
francois.djindjian@wanadoo.fr

Résumé

Dans cette introduction au thème des mégadonnées en archéologie, l'accent est mis sur le concept relatif de mégadonnées dans le temps de l'évolution technologique de l'informatique et dans l'histoire récente de l'archéologie computationnelle.

Mots-clés : Mégadonnées, archéologie, informatique

Abstract

In this introduction to the topic of Big Data in archaeology, emphasis is placed on the concept of Big Data in the time of the technological evolution of computing and in the recent history of computational archaeology.

Keywords: Big Data, archaeology, computer

1. Introduction

D'après Wikipedia (article Big Data), le volume des données stockées au niveau mondial est en pleine expansion : les données numériques créées dans le monde seraient passées de 1,2 zettaoctet par an en 2010 à 2,8 zettaoctets en 2012 et s'élèveront à 40 zettaoctets en 2020. À titre d'exemple, Twitter générait en janvier 2013, 7 téraoctets de données chaque jour et Facebook 10 téraoctets. Ce sont les installations technico-scientifiques (météorologie, astronomie, CERN, etc.) qui produiraient le plus de données. Le radiotélescope « Square Kilometre Array » par exemple produira 50 téraoctets de données analysées par jour, tirées de données brutes produites à un rythme de 7000 téraoctets par seconde.

Le volume des données produit par l'archéologue n'est pas évidement du même ordre de grandeur. Un chantier de fouilles archéologiques produira, après plus d'une dizaine d'années de campagnes annuelles, des données de l'ordre d'une ou plusieurs centaines de Moctets, compatibles avec les capacités de stockage des ordinateurs de bureau. Ces volumes sont dus pour l'essentiel à la numérisation des relevés et des documents photographiques. Mais, de plus en plus, sont utilisées en archéologie des applications grosses productrices de volume de données comme le 3D, le Lidar ou les analyses de laboratoire.

Par ailleurs, et indépendamment, se pose la question de l'archivage des données, dans un contexte institutionnel où les données archéologiques restent sous la responsabilité directe de l'archéologue, dans son environnement informatique individuel, et dont la sécurité de l'archivage n'est donc ni sûr ni pérenne.

2. Historique du concept de Big Data

A ses origines, l'archéologie a été une science de l'objet et les archéologues, eux-mêmes, se désignaient souvent comme antiquaires ou collectionneurs.

C'est à partir des années 1960, que l'archéologie devient progressivement une science de l'information des sociétés du passé : informations intrinsèques qui décrivent les artefacts de la culture matérielle et informations extrinsèques qui enregistrent le contexte de ces artefacts et leurs relations.

Ces informations étaient disséminées par le support écrit : livres, corpus, monographies, articles dans les revues académiques, consultables dans les bibliothèques institutionnelles et privées.

Les archéologues archivaient leurs documents de travail : carnets de fouilles, relevés stratigraphiques et planigraphiques, dessins d'objets, plans, photographies, inventaires, mensurations, mesures, notes, projets d'articles, tirés à part ainsi que les communications épistolaires entre archéologues dans le milieu académique. Tous ces documents faisaient l'objet au mieux d'un archivage institutionnel ou privé.

Le développement de l'informatique transforma progressivement le support papier en support électronique : systèmes bibliographiques de l'Information scientifique et technique, systèmes documentaires (« banques de données »), fichiers d'inventaires et de mesures.

La machine à écrire (comme la célèbre IBM à boule ou la petite portative), apparue dans la seconde moitié du XIX° siècle, disparait à la fin des années 1980 (tout comme le métier de dactylo), remplacée par le microordinateur et le logiciel de traitement de texte.

Le courrier devient messagerie mais le message n'est plus archivé sauf exception. L'historiographie perd ainsi les échanges privés entre chercheurs, souvent plus instructifs que les échanges officiels.

Les tirés à part d'articles publiés, ou leur photocopie, sont remplacés par des fichiers en format pdf. Ils sont échangés ou accessibles sur des sites en ligne en accès libre ou en vente sur les sites des éditeurs privés.

Le dessin, activité manuelle (les laboratoires employaient des ITA dessinateurs), devient dessin assisté par ordinateur (DAO) avec la fameuse suite Adobe : dessin vectoriel (Illustrator), composition (Pagemaker/Indesign), création et retouche d'image (Photoshop) et ses équivalents concurrents.

Puis, à partir des années 1990, la numérisation s'accéléra, qui engendra pour l'archéologie de nombreuses nouvelles données :

- les mesures physico-chimiques,
- la prospection géophysique (terrestre et maritime),
- Les données Lidar,
- la cartographie,
- le Système d'Information Géographique (SIG),
- la photographie numérique,
- la numérisation des photographies argentiques,
- le film numérique,
- la numérisations des relevés stratigraphiques et planigraphiques,
- et enfin le 3D avec la réalité virtuelle et la photogrammétrie numérique.

Dès lors, la question du Big Data s'imposa à tous.

3. Big Data : une longue histoire liée aux progrès de l'informatique

Le concept de « Big Data » est relatif. Il est lié aux problèmes que posent l'archivage et le traitement de grands volumes de données en rapport avec la disponibilité du hardware (stockage des données)

et des outils logiciels pour les rechercher (systèmes documentaires), les consulter, en extraire une partie, les visualiser (systèmes graphiques, SIG, 3D) et les traiter (visualisation graphique, analyse des données multidimensionnelles, modélisation, intelligence artificielle, etc.).

Le monde scientifique moderne aime à ressusciter les problématiques échouées face aux difficultés technologiques du moment par des nouveaux noms désignant les mêmes concepts. L'intelligence artificielle, ce grand mythe des temps modernes, en est un bon exemple : issue de la cybernétique d'avant-guerre, elle est née dans les années 1950 (perceptron de Rosenblatt) avec les premiers ordinateurs, et, elle se relance périodiquement sous différents noms : IA, apprentissage automatique, système expert, réseau de neurones, moteur de règles, et dernier en date, apprentissage profond. Ses applications les plus réussies se retrouvent dans la robotique, la traduction automatique, la reconnaissance de formes, l'aide au diagnostic, l'aide à la décision, le traitement des Big Data (où elle remplace le data mining des années 1990) et la plus médiatisée de toutes, les jeux (quand la machine bat l'humain : échecs, Go).

Le Big Data aussi a déjà une longue histoire. Il est lié à l'évolution de la taille des mémoires des ordinateurs et au volume de stockage des mémoires de masse (disques et bandes magnétiques). Dans les années 1960/70, les mémoires (à tores de ferrites) étaient limitées à plusieurs dizaines ou centaines de kilooctets. Les mémoires RAM actuelles (des circuits imprimés) font plusieurs à plusieurs dizaines de Gigaoctets, soit un million de fois plus ! Le stockage sur mémoire de masse a connu la même évolution technologique depuis les 2 Megaoctets du premier disque dur d'IBM en 1962, les 300 Megaoctets dans les années 1980, les 25 Gigaoctets en 1998 et plusieurs Teraoctets actuellement soit un million de fois plus aussi !

Les bandes magnétiques, organisées en baies de stockage qui peuvent contenir une dizaine ou une vingtaine de bandes magnétiques, peuvent atteindre une capacité totale jusqu'à plusieurs dizaines de téraoctets. Les bibliothèques de bandes sont donc le moyen le plus aisé d'assurer la sauvegarde et l'archivage de données volumineuses, comme pour les grandes fermes informatiques du Web ou le stockage institutionnel des organismes de recherches. Hélas la durée de vie d'une bande magnétique n'est que d'une vingtaine d'années !

Dans les années 1970, les banques de données (systèmes documentaires) et les grands tableaux étaient les « Big Data » de cette période. En France, c'est la grande période institutionnelle des systèmes documentaires mis en œuvre par le ministère de la Culture (musées, Inventaire général des monuments et richesses artistiques de la France, Carte archéologique), utilisant le logiciel Mistral de Bull. Mais les données de cette période sont du texte, les images étant stockées sur microfiche et consultables sur un lecteur installé à côté du terminal. C'est seulement à partir des années 1980, que les progrès technologiques des mémoires, des unités de stockage (disques magnétiques, vidéodisque puis disque optique numérique) et des réseaux ont vus arriver les premiers prototypes de serveur données/images/son qui deviennent opérationnels dans les années 1995, avec le développement d'Internet. Notons cependant que le système vidéotex, le précurseur d'Internet, a été opérationnel en France à partir de 1980 jusqu'en 2012.

Les grands tableaux, qui sont les données de base des archéologues pour la plupart des problématiques traitées (Djindjian 1991, 2011), faisaient l'objet de manipulations graphiques dans les années 1960, avant d'être traités par analyse des données multidimensionnelles dans les années 1975, malgré les limitations des ordinateurs en puissance de calcul et en mémoire centrale. A partir des années 1990, ces limitations ont disparues et ces traitements ont commencé à être effectués sur microordinateur.

Les années 1980 et 1990, qui sont les années du développement du microordinateur, des réseaux et des logiciels bureautique, voient l'archéologue s'approprier individuellement ces outils et l'institution se trouve alors en retrait sur des projets communautaires.

Les années 1990 voient l'arrivée d'un nouveau vocabulaire sinon d'une nouvelle approche, le Data mining (ou exploration de données) qui applique les techniques statistiques multidimensionnelles à de grand corpus de données comme ceux obtenus par les habitudes de consommation, de consultation de données sur Internet ou de questionnaires et qui permettent d'identifier des types de comportements de consommateurs (segmentation, scoring). Les techniques d'apprentissage font également leur apparition. Mais les méthodes de l'archéologie ne se sentent pas concernées par les intérêts essentiellement marketing du data mining.

Les années 2000 voient l'émergence du vocabulaire des Big Data (en français, mégadonnées), liée à la production massive (« orwellienne ») des données que le progrès technologique de l'informatique permet aujourd'hui de stocker, de communiquer par des réseaux, de visualiser et de traiter. Les organismes institutionnelles de la recherche commencent à s'émouvoir de la dispersion des données enregistrées par les chercheurs individuels (mais financés par l'institution) et qui se perdent quand le micro-ordinateur tombe en panne ou quand le chercheur part à la retraite, et tout particulièrement dans le domaine des Sciences humaines et sociales où le chercheur individuel prime sur le laboratoire.

En France, le CNRS lance le projet TGIR Huma-Num (www.huma-num.fr) du CNRS pour l'archivage des données numériques des Sciences humaines. Il s'agit d'une plateforme informatique permettant l'acquisition, le stockage, la dissémination, le traitement et l'archivage des données. Plusieurs laboratoires d'archéologie se sont regroupés au sein du consortium Masa (Mémoire des archéologues et des sites archéologiques) pour utiliser les services du TGIR Huma-Num. Il a pour objectif de proposer un accès unifié à des corpus variés de données et de documentations produites par les archéologues. Il développe des méthodes et des outils à destination de la communauté archéologique, en respectant les standards internationaux (https://masa.hypotheses.org/).

Au niveau européen, le projet Ariadne a lancé des coopérations entre les archéologies européennes sur des projets fédérateurs (notamment les thésaurus) et des plateformes de service, dont le sujet de l'archivage (https://ariadne-infrastructure.eu/).

4. Quelles mégadonnées archéologiques ?

Les mégadonnées archéologiques sont constituées d'un ensemble non limitatif de fichiers de taille, de format et de structure variable :

- Des bases de données, résultats de l'enregistrement des données de fouilles archéologiques (informations extrinsèques) et de la description des artefacts (informations intrinsèques). Ces données sont enregistrées dans des logiciels variés depuis les logiciels de traitements de texte, les tableurs jusqu'aux systèmes de gestion de bases de données.
- Des textes anciens dans leur écriture d'origine et leur traduction (philologie),
- Des banques de données créées avec des logiciels de recherche documentaire,
- Des documents numérisés : photos numériques, diapositives numérisées, relevés stratigraphiques et planigraphiques numérisés, films vidéo numérisés, 3D,
- Des documents graphiques vectoriels comme ceux créés par des logiciels de PAO ou des logiciels de système d'information géographique (SIG),
- Des tableaux de données quantitatives,
- Des fichiers de mesures comme ceux produits par des appareillages physico-chimiques : prospection géophysique, Lidar, spectrométrie variée, datations, etc.

5. Les fonctions d'un service « mégadonnées »

Les fonctions d'un service de mégadonnées ne sont pas limitées à l'archivage. Elles concernent l'ensemble de la chaine depuis l'acquisition (*Submission information package*), le stockage, la

signalisation (c'est-à-dire l'indexation ainsi que la définition et la gestion des métadonnées qui décrivent les données), la diffusion (qui permet la consultation par Internet), l'archivage (suivant un format standardisé), la sélection (qui permet d'extraire et de formater les données pour un traitement) et le traitement.

La fonction de traitement est riche et variée et comprend tous les outils et logiciels utilisés depuis plus de cinquante ans : analyse lexicographique, statistiques, analyse des données multidimensionnelles, système d'information géographique, traitement d'image, modélisation, 3D, et plus récemment le retour de l'intelligence artificielle utilisant les techniques d'apprentissage automatique (deep learning) etc.

6. Les bonnes pratiques

Au-delà du plaisir de s'enivrer de mots à la mode, l'archéologue doit s'investir dans le domaine des projets, qui mêlent efficacement nouveautés techniques et pragmatisme. Les bonnes pratiques sont alors le meilleur garant d'un projet réussi.

Les métadonnées, qui sont les données qui décrivent les données regroupent deux ensemble : les métadonnées individuelles liées au données produites par l'archéologue et les métadonnées communes, institutionnelles, globales et spécialisées, de plus en plus normalisées. Ces métadonnées institutionnelles sont issues en archéologie des projets documentaires des années 1970 (Ministère de la Culture, CNRS, Information scientifique et technique (INIST)) qui ont investi dans la réalisation des premiers grands thésaurus, qui sont les bases des métadonnées actuelles. En archéologie, le thésaurus de référence est Pactols développé à l'origine pour le projet signalétique Frantiq à la Maison de l'Orient Méditerranéen, qui possède 30 000 références (conforme à la norme ISO 25964 des thésaurus multilingues). Les thésaurus du Ministère de la Culture (Inventaire général, banques de données muséographiques) ont été regroupées sur la plate-forme Ginco.

Les normes qui homogénéisent les productions industrielles depuis plus de cinquante ans, concernent également progressivement l'archéologie, soit indirectement par des logiciels génériques, soit directement mais encore rarement par des normes dédiées à l'archéologie.

L'archivage (OAIS, Open Archival Information System) possède son propre standard, l'ISO 14721:2012. Dans cette norme, un « *paquet d'information* » contient les informations à archiver, à conserver ou à communiquer aux utilisateurs. Le paquet d'information contient toujours l'objet que l'on veut conserver, et les métadonnées nécessaires à sa préservation. Trois types sont définis :

- Le paquet d'information à verser (SIP): Produit par de dépositaire de l'archive, selon le modèle imposé par le gestionnaire de dépôt ;
- Le paquet d'information archivé (AIP): Contenus (*Content Data Objects*) et métadonnées. Produit par et pour le gestionnaire de dépôt ;
- Le paquet d'information diffusé (DIP): en fonction des droits de l'utilisateur qui effectue la requête et des droits de diffusion.

La norme CIDOC-CRM (ISO 21127:2014) est une norme propre au patrimoine culturel et à ce titre elle est concernée par le thème des Big Data et de l'archivage.

Les données élémentaires

Le paquet d'information à verser doit contenir les informations au niveau le plus élémentaire connu. Le système d'archivage doit posséder les fonctions de sélection, de filtrage et d'agrégation utiles à construire les données à n'importe quel un niveau d'agrégation supérieur. Dans le cas contraire, les informations au niveau le plus élémentaire sont définitivement perdues.

Les données brutes

Le paquet d'information à verser doit contenir des données brutes (raw), avec la définition la meilleure possible, sans format ou traitement visant à diminuer le volume ou à modifier la donnée.

7. Les traitements

Il est illusoire de penser qu'une accumulation de données puisse sous l'action de quelques algorithmes, aussi puissants soient-ils, être capable, spontanément, de fournir des résultats : des connaissances ou des décisions.

L'exploration des données (c'est le nom donné aux différentes méthodes que l'on désignait dans les années 1970 sous le nom d'analyse des données multidimensionnelles) ne peut être efficace que dans le cadre d'une construction formelle qui puisse permettre à la fois de mettre en évidence une structure dans les données et de pouvoir la valider.

C'est sans doute cette trop grande confiance (ou une trop grande paresse) qui est à l'origine du désappointement dans l'utilisation de ces techniques, qui sont en retrait à partir des années 2000 avec le succès du post-modernisme.

L'intégration des techniques d'exploration des données dans un processus cognitif global nécessite une approche à plusieurs niveaux, comme celle que nous avions proposée, sous le titre de « le triplet systémique » (Djindjian 2002) à partir des résultatsde notre thèse (Djindjian 1980, note 1). Ces processus, pour être réellement cognitifs, doivent intégrer explicitement des mécanismes d'apprentissage, que l'analyse des données réalise par le jeu des éléments supplémentaires et par l'itération sur les informations intrinsèques permettant une interaction archéologue-objet, mécanisme de l'apprentissage.

L'enjeu de l'intelligence artificielle, à travers les différents algorithmes qu'elle a développés depuis les années 1950, peut en fait se résumer dans le paradoxe suivant : utiliser la puissance de calcul toujours croissante des ordinateurs avec des algorithmes itératifs simples ou mettre en œuvre une construction formelle sophistiquée. L'analogie avec le jeu d'échec illustre bien ce paradoxe : soit calculer toutes les combinaisons possibles soit concevoir une stratégie de jeu qui réduit le calcul des combinaisons. La première option, dont le succès n'est dû qu'à l'amélioration des capacités de calcul de l'ordinateur, n'est qu'une étape préparant la seconde option, d'où le succès du terme et sans doute à terme des résultats du concept d'apprentissage profond, qui doit dépasser le fait d'être le mot à la mode.

8. Conclusions

Au-delà du terme « mégadonnées », se trouve en fait le rapport entre le chercheur scientifique et l'évolution fantastique de la technologie des ordinateurs dans la deuxième moitié du XX° siècle. Plus cette technologie offre des moyens supplémentaires (capacité de calcul, volume de stockage, canal de communication), plus les besoins apparaissent (souvent plus avec l'aide d'un bon marketing des industriels qu'avec une attente exprimée des chercheurs). L'archéologie a suivi cette tendance avec des besoins certes incomparablement plus faibles, mais le développement de certaines méthodes (comme le 3D) comme la sociologie particulière de l'archéologue implique que l'institution se mobilise pour offrir des environnements, des standards et des services pour les mégadonnées archéologiques.

References

Djindjian, F. 1980. *Constructions de systèmes d'aides à la connaissance en archéologie préhistorique. Structuration et affectation: méthodes et algorithmes.* 2 volumes. Thèse 3eme cycle, Archéologie préhistorique : Paris I.

Djindjian, F. and Ducasse, H. (eds) 1987. *Mathématiques et informatique appliquées à l'archéologie*. Conseil de l'Europe : PACT 16.
Djindjian, F. 1991. *Méthodes pour l'Archéologie*. Paris : Armand Colin.
Djindjian, F. 2002. Pour une théorie générale de la connaissance en archéologie, in *XIV Congrès International UISPP, Liège Septembre 2001. Colloque 1.3. Archeologia e Calcolatori* 13: 101-117.
Djindjian, F. 2011. *Manuel d'Archéologie*. Paris : Armand Colin.

Note 1

Définition du triplet systémique S (O, I, E): les objets O, l'information intrinsèque I et l'information extrinsèque E.

- Etape 1: Définition du système S :
  Le système S est lui-même défini, par un ensemble de valeurs constantes de E, comme par exemple les objets d'une même unité stratigraphique (ensemble clos), d'une même sépulture, les peintures d'une même grotte ornée, les outils d'une même structure d'habitat, les structures urbaines contemporaines d'un même territoire, etc. qui peuvent toutes être définies par un jeu de valeurs constantes d'informations extrinsèques de type T (temps), H (structure d'habitat), R (territoire), L (localisation), M (origine), EV (environnement), EC (économie), etc.

- Etape 2: Perception et description des informations intrinsèques I,

- Etape 3: Enregistrement des informations extrinsèques E,

- Etape 4: Formalisation de la structuration:
  Structurer le système formalisé par le tableau Objets x Description des objets (O x I), qui fournit des structures de partitions (classifications ou typologies) ou des structures sérielles (sériations), donnant un nouvel ordre sur O, soit O+, et des corrélations sur I, soit I+. Le système passe alors de l'état cognitif S (O, I) à l'état S+ (O+, I+). Cette structuration est appelée structuration intrinsèque.
  Structurer le système formalisé par le tableau d'occurrence (I x E), qui fournit des structures de correspondances entre les deux ensembles d'informations, structuration en faciès chronologiques pour E=T, structuration spatiale pour E=H ou L, déterminisme environnemental pour E = Ev, etc. Le système passe alors d'un état cognitif S (O, I, E) à un état cognitif S+ (O+, I+, E+). Cette structuration est appelée structuration extrinsèque.

- Etape 5: Application des techniques d'analyse des données sur les tableaux (O x I) ou (I x E),

- Etape 6: Rétroactions par retour sur I et E (c'est un mécanisme d'apprentissage),

- Etape 7: Enrichissements progressifs par intégration de nouveaux I et E ,

- Etape 8: Validation (sur un autre système de O, par une autre corrélation E, etc.).