DataLAC : intelligence artificielle dans un lac de données pour valoriser les archives archéologiques

Rajae El-Idrissi*, Elias Makhlouf*, Mohamed Riad Sahrane*, Dewanou Romeo Elohounkpon*, Jorje Antonion Sanchez Cristancho**, Josefina Simon Reig**, Laura Romero**, Juba Agoun*, Jean-Pierre Girard***, Gabriel de Prado**, Jérôme Darmont*, Sabine Loudcher*

*Université Lumière Lyon 2, ERIC, Lyon, France
**Musée d'Archéologie de Catalogne, Site d'Ullastret, Espagne
***Université Lumière Lyon 2, Archéorient, Lyon, France

Résumé. Notre travail dans le cadre du projet DataLAC explore comment l'extraction automatisée de métadonnées depuis des carnets de fouille manuscrits, au sein d'un lac de données intégrant des sources de données hétérogènes, peut transformer l'exploitation d'archives archéologiques. En s'appuyant sur le cas du site ibérique d'Ullastret en Catalogne et ses trente années de documentation de terrain (1947-1977), le projet vise à démontrer l'apport d'un lac de données dont l'architecture comprend quatre phases : (a) numérisation des carnets manuscrits et de la documentation scientifique ; (b) transcription automatique par apprentissage automatique ; (c) extraction et modélisation des métadonnées enrichies sémantiquement ; (d) interrogation sémantique des documents et métadonnées. L'enjeu central réside dans la capacité du lac de données à réunir des sources documentaires dispersées et à en extraire automatiquement des informations structurées permettant de nouvelles formes d'analyse. DataLAC constitue ainsi une preuve de concept sur la manière dont l'intelligence artificielle appliquée à l'extraction de métadonnées peut renouveler les pratiques de recherche en archéologie, en rendant exploitables et interrogeables des corpus jusqu'alors difficiles d'accès.

1 Introduction

L'archéologie est une discipline qui ne peut jamais reproduire la source scientifique de son savoir, la fouille archéologique, et n'en n'étudie que des représentations, aujourd'hui massivement numériques, mais naguère manuscrites. Elle repose aussi sur les rapprochements de vestiges comparables dans leurs caractéristiques morphologiques (photographies et représentations publiées, croquis illustrant des carnets de fouille manuscrits ou relevés à main levée sur le terrain). L'automatisation de l'analyse de ces corpus est un enjeu et repose sur une reconnaissance graphique et sémantique de ces archives scientifiques extrêmement diversifiées.

Le projet DataLAC (données, archives, et textes archéologiques : création et exploitation d'un lac de données sémantiques pour l'archéologie de la Catalogne) vise à relever ce défi sur un site ibérique de la première moitié du VI e siècle av. J.C. , le site d'Ullastret en Catalogne

(Espagne). Ce site est fouillé de façon continue depuis 1947 mais les documents des trente premières années de fouille n'ont pas été numérisés et encore moins transcrits en « *texte numérique* ». Ces documents (cahiers de fouille, archives photographiques, inventaires des objets découverts) très hétérogènes posent d'importants défis en matière de numérisation, de reconnaissance d'écriture manuscrite et de mise en relation sémantique.

Pour maîtriser l'hétérogénéité de la documentation du site, pour gérer la qualité et faciliter le partage et l'analyse, le projet DataLAC mobilise le concept de lac de données qui vise à conserver ces dernières, qu'elles soient primaires ou déjà enrichies, dans leur forme/format originel(le). Elles demeurent ainsi, à condition d'être soigneusement documentées par des métadonnées, des références mobilisables. La modélisation des métadonnées décrivant les entités de données d'un lac de données est cruciale et elle doit être assortie d'un système de gestion des métadonnées efficace pour permettre une interrogation et une analyse efficaces des données hétérogènes et de leurs métadonnées.

En d'autres termes, le projet DataLAC vise à développer une preuve de concept à la fois opérationnelle et reproductible, permettant d'exploiter et d'analyser des corpus archéologiques en mobilisant des outils de science des données et d'intelligence artificielle au sein d'un lac de données intégrant sources hétérogènes et métadonnées. Pour atteindre son objectif le projet comprend quatre phases :

1. L'acquisition des données avec la numérisation des treize cahiers de fouilles manuscrits (notes de terrain) et de la documentation scientifique (rapports, photos, diapositives, plans, etc.) en haute résolution, suivie d'un prétraitement des images afin d'obtenir une qualité adaptée aux automatisations ultérieures. L'ensemble de ces données constitue le corpus de documents du projet.
2. La transcription automatique des 4000 pages des treize carnets de fouilles grâce à la reconnaissance automatique d'écriture manuscrite (ou en anglais *Handwritten Text Recognition* HTR) à l'aide de modèles d'apprentissage automatique affinés.
3. La modélisation et l'enrichissement des métadonnées décrivant les carnets de fouille et la documentation scientifique. Chaque carnet, page de carnet, ligne de carnet, document scientifique est décrit par plusieurs métadonnées. Ces dernières sont mises en relation grâce à un thésaurus archéologique multilingue permettant de produire des métadonnées normalisées et interrogeables.
4. L'interrogation sémantique. Les documents et leurs métadonnées sont organisés dans un lac de données et sont gérées accessibles via un système de gestion des métadonnées composé d'une base de données relationnelle, d'une API RESTful et d'une interface Web pour faciliter la recherche, la visualisation et l'annotation des documents du corpus.

La suite de l'article présente chacune de ces phases.

2 Corpus et thésaurus

Le corpus du projet DataLAC rassemble des archives qui retracent trois décennies de fouilles archéologiques du village ibérique de Puig de Sant Andreu à Ullastret[1]. Au cœur

1. http://www.macullastret.cat/

de cette collection se trouvent treize carnets de terrain manuscrits, rédigés entre 1947 et 1977, en catalan et en castillan, par l'archéologue Miquel Oliva i Prat et ses collaborateurs. Leur contenu se distingue par une grande hétérogénéité : quelques 4000 pages mêlent observations narratives, annotations techniques, croquis d'artefacts ou de fouilles, diagrammes stratigraphiques, plans de fouilles et notes de terrain provisoires. À ces carnets s'ajoutent de nombreux feuillets complémentaires, manuscrits ou dactylographiés, insérés au fil des carnets. Ces documents additionnels incluent, des photographies de terrain ou encore des matériaux éphémères produits au cours des campagnes de fouilles.

En plus des carnets, le corpus comporte également un vaste ensemble de documents graphiques et de supports visuels, tels des dessins techniques (relevés de terrain, planches de profils de céramiques), des photographies techniques (vestiges sur le terrain, couches stratigraphiques, planches d'objets), des cartes et plans de diverses époques. Tous ces documents représentent les excavations, les artefacts, les contextes de terrain et les archéologues au travail. L'ensemble a été numérisé par le musée-site d'Ullastret et mis à disposition du projet.

Chaque carnet, page de carnet, croquis, feuillet supplémentaire, photo, plan, etc. est décrit par un ensemble de métadonnées. Une première catégorie de métadonnées est directement extraite des signatures techniques embarquées dans les fichiers résultant de la numérisation ; ces métadonnées primaires ne suffisent pas, elles doivent être enrichies de sémantique permettant de décrire finement le contenu des carnets et des documents scientifiques (« *ce document représente cet artefact* » ou « *cet artefact est dans la zone A fouillée par Oliva, zone qui correspond aux zones modernes X et Y* ») ou permettant de faire également des liens entre tous les documents du corpus (« *ces deux documents décrivent le même artefact* »). Le projet DataLAC s'est doté d'un ensemble de descripteurs terminologiques pour pouvoir constituer des métadonnées riches sémantiquement. L'ensemble de ces descripteurs terminologiques est organisé dans un thésaurus[2] spécialisé bilingue (catalan-castillan, qui sera ultérieurement apparié avec des concepts en français). Avec une granularité très fine, il est fondé sur le vocabulaire utilisé dans les carnets de fouille. Ce thésaurus constitue un outil puissant d'une part pour l'annotation sémantique et pour la normalisation terminologique des descripteurs, d'autre part par les relations logiques qu'il introduit entre les labels différents de concepts identiques ou proches. A terme il sera aligné sur les références archéologiques, contribuant ainsi à l'interopérabilité du corpus du projet.

3 Segmentation et transcriptions automatiques des carnets

La transcription automatique des pages des carnets de fouille comporte deux étapes : (1) extraction et structuration des éléments graphiques pour reconnaître les positions des lignes de texte, des illustrations (*e.g.,* des croquis) et des zones vierges dans les pages des carnets ; (2) production d'une transcription exploitable des lignes de texte (*voir Fig.* 1). Ce processus de reconnaissance de l'écriture manuscrite (ou en anglais *Handwriten Text Recognition*, HTR) s'appuie sur des méthodes d'apprentissage automatique adaptées à la nature hétérogène du corpus.

Toutes les opérations de segmentation et de transcription sont effectuées à l'aide de l'application `eScriptorium`[3] et du moteur d'HTR `Kraken`, installés localement. Deux modèles

2. `https://thesaurus.mom.fr/?idt=th62`
3. `https://escriptorium.rich.ru.nl/` et `https://gitlab.com/scripta/escriptorium/`

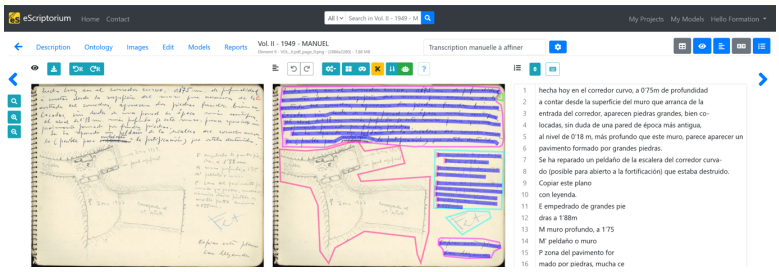

FIG. 1 – *Segmentation et transcription des pages des carnets de fouille*

pré-entraînés ont servi de point de départ avec un spécialisé dans la segmentation des zones de texte et des lignes, et un autre pour la transcription proprement dite des zones segmentées. Pour garantir la compatibilité et l'interopérabilité des données produites, les résultats sont exportés au format XML ALTO, avec des balises conformes aux principes de l'ontologie `SegmOnto`[4] (Gabay et al., 2024).

3.1 Segmentation des pages

Le modèle de segmentation `blla.mlmodel` de (Kiessling, 2020) est utilisé afin d'identifier automatiquement les zones textuelles et les lignes de texte dans les pages des carnets. Ce modèle[5], à base de réseaux de neurones convolutifs (CNN), est le modèle de segmentation de `Kraken` pré-entrainé à l'aide de toutes les données de Gallicorpora[6].

Appliqué sur les pages des carnets de fouille, le modèle n'arrive pas à distinguer correctement les zones non textuelles, en particulier les illustrations (croquis, photos collées, plan manuscrit) et les zones vierges dans les pages. Une correction manuelle est nécessaire pour ajuster les résultats de segmentation aux besoins spécifiques du projet, à savoir l'ajout de zones graphiques et de zones vides comme illustré dans la figure 1 . Ces corrections, réalisées sur les zones segmentées par le modèle des quatre premiers carnets, permettent de constituer un jeu de données d'entraînement de 364 pages visant à affiner le modèle de segmentation `blla.mlmodel`. L'entraînement, réalisé sur un GPU, a nécessité environ 07 heures de calcul. Estimée par validation croisée, la proportion de pixels correctement classés est de 97,2 %.

Pour garantir une conformité maximale aux standards d'annotation adoptés, le tag Text (présent dans le modèle initial) est fusionné avec le tag MainZone afin d'harmoniser la sortie du modèle avec la structure exigée par l'ontologie `SegmOnto`. Durant la phase d'entraînement, nous avons utilisé l'option `resize new` pour ajuster dynamiquement la couche de sortie du modèle et ne conserver que les tags présents dans les données d'entraînement, éliminant ainsi les tags non représentés dans le corpus. Ceci améliore la cohérence du modèle et réduit le bruit lors de la prédiction. Le score moyen d'intersection (*Intersection over Union* IoU) s'est stabilisé à 36,6%, une valeur relativement faible, expliquée par la formation de

4. `https://segmonto.github.io/`
5. `https://zenodo.org/records/14602569`
6. `https://github.com/Gallicorpora/Segmentation-and-HTR-Models`

contours flous autour des zones. En effet, les carnets de terrain de l'archéologue contiennent de petites lignes de quadrillage qui introduisent probablement du bruit dans le processus de détection des zones. Néanmoins, le modèle conserve une bonne performance globale et assigne correctement les nouveaux labels aux zones détectées. L'IoU pondéré atteint 79%, indiquant une bonne performance globale du modèle, notamment pour les classes principales telles que les MainZone .

Le modèle affiné est ensuite appliqué sur le reste des carnets pour segmenter toutes leurs pages puis les erreurs produites sont corrigées manuellement. Même si le modèle affiné produit encore des erreurs, son utilisation (comparée à celle du modèle non affiné) permet de réduire quand même significativement le temps nécessaire pour corriger les positions des différentes zones segmentées.

3.2 Transcription des pages

La phase de transcription automatique débute par l'application du modèle `Manu McFrench`[7], créé par (Chagué et al., 2023), à toutes les zones de texte segmentées. Ce modèle a été retenu car il a été pré-entraîné sur un corpus de manuscrits écrits en français, en espagnol ou en anglais, sachant que peu de modèles de transcription sont pré-entraînés sur des manuscrits écrits en espagnol et encore moins en catalan. Sur la base de 228 pages transcrites manuellement des trois premiers carnets (le reste des pages de ces carnets étant des pages blanches), le modèle `Manu McFrench` est affiné. L'option `resize new`, option spécifique à `Kraken`, est activée pendant l'entraînement pour ajuster dynamiquement la couche de sortie du modèle et n'inclure uniquement les caractères présents dans les données d'entraînement, éliminant ainsi les glyphes non représentés dans le corpus. Cela améliore la cohérence typographique du modèle et réduit le bruit lors de la prédiction. L'évaluation du modèle affiné a donné des résultats globalement satisfaisants. Le taux moyen de reconnaissance des caractères, estimé par validation croisée, a atteint 90,6%, tandis que le taux de reconnaissance des mots, plus strict, s'est établi à 67,8%. Cette différence, attendue lors du traitement de textes manuscrits, est due à la sensibilité accrue de la seconde métrique aux erreurs ponctuelles : une seule lettre incorrecte peut invalider un mot entier. Malgré ces limites, les performances obtenues confirment l'adéquation du modèle pour une application à grande échelle. Son utilisation sur les carnets restants accélère significativement la transcription des pages, mais sur une phase de correction a posteriori reste nécessaire. Il est important de rappeler que l'objectif n'est pas seulement d'obtenir les meilleures performances possibles, mais de produire une transcription complète de toutes les pages des carnets. Il a fallu trouver un équilibre entre le temps consacré à la transcription manuelle de premiers carnets, celui à l'amélioration des modèles et celui nécessaire à la correction des erreurs. Ce processus a réduit de manière significative le temps requis pour la transcription et la localisation des différentes zones d'intérêt, une tâche qui aurait été beaucoup plus chronophage si elle avait été réalisée entièrement à la main.

4 Lac de données sémantiques

Pour lever les verrous informatiques liés au stockage, à l'interrogation, à l'analyse et à la visualisation des données hétérogènes, le projet DataLAC utilise le concept de lac de données

7. `https://zenodo.org/record/6657809`

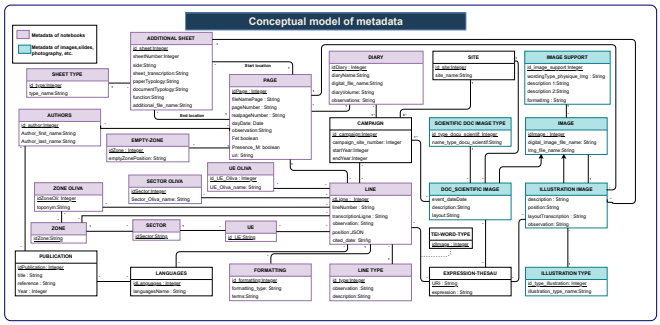

FIG. 2 – *Modèle conceptuel des métadonnées*

qui propose de stocker les données dans leur format d'origine, sans schéma prédéfini et sans transformation préalable à leur stockage (Dixon, 2010; Hai et al., 2016). Avec un tel principe, tous types de données peuvent cohabiter dans un lac de données, qu'elles soient structurées ou non. Toutefois, pour être exploitable, un lac de données a besoin de métadonnées qui permettent de décrire et de relier les entités de données stockées dans le lac, ainsi que d'un système efficace de gestion de ces métadonnées.

4.1 Système de gestion des métadonnées

Dans des travaux antérieurs au projet DataLAC, nous a étendu la définition du concept de lac de données, ainsi que les fonctionnalités qu'un système de gestion de métadonnées doit avoir pour être complet et efficace (Sawadogo et al., 2019). De plus, nous avons proposé un méta-modèle (Scholly et al., 2021) de métadonnées qui comprend trois niveaux de modélisation (conceptuel, logique et physique). La figure 2 illustre le modèle conceptuel des métadonnées décrivant les carnets, leurs pages et les lignes, les illustrations et feuilles supplémentaires présentes dans les carnets ainsi que toute la documentation scientifique. Les mots-clés, désignés par *expression-thesau* dans la figure 2 et piochés dans le thésaurus, permettent de décrire sémantiquement tous les éléments. Les zones, secteurs et unités stratigraphiques d'Oliva permettant de localiser les descriptions faites dans les lignes des pages des carnets sont également dans le thésaurus. Ainsi toutes les entités du corpus sont décrites par une série de métadonnées et le thésaurus permet de les relier entre elles.

4.2 Architecture du lac de données

Le système de gestion des métadonnées comprend une interface Web conçue pour faciliter la recherche, la visualisation et l'annotation des documents du corpus (*cf.* FIG. 3). Elle permet aux archéologues la saisie, la modification, la consultation, l'interrogation et la mise en relation des métadonnées. L'interface communique avec le back-end via une API RESTful. Au niveau du back-end, les métadonnées sont gérées dans une base de données relationnelle PostgreSQL. Le thésaurus est intégré dans le système de gestion de métadonnées ; il est hébergé

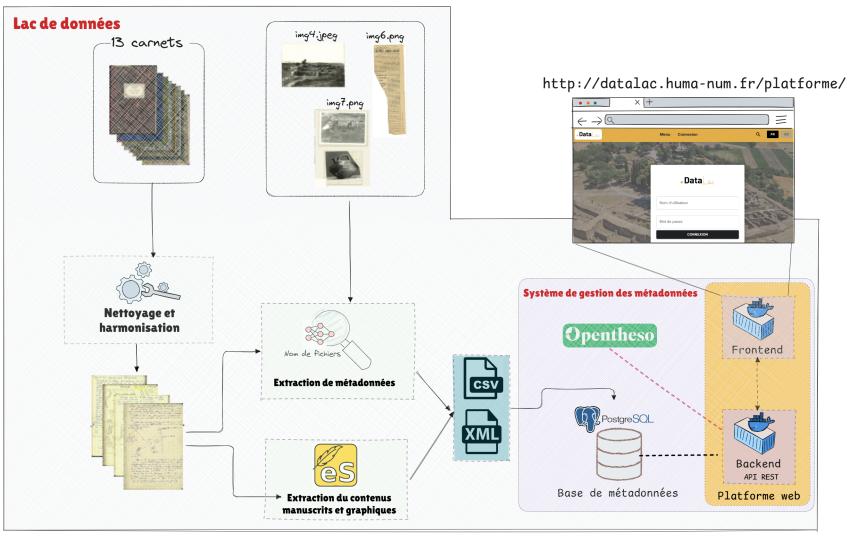

FIG. 3 – *Architecture du lac de données*

sur la plateforme Opentheso[8] de la Maison de l'Orient et de la Méditerranée. Cette architecture modulaire permet non seulement d'avoir un système efficace de gestion des métadonnées, mais elle soutient également de futures évolutions, telles que l'intégration de nouveaux types de documents, l'alignement du thésaurus sur des référentiels archéologiques et le développement d'outils analytiques avancés. Pour assister les archéologues dans la description de chaque document du corpus, tâche fastidieuse et chronophage, des scripts Python extraient des noms de fichiers les métadonnées de base et puis ces dernières sont insérées automatiquement dans la base des métadonnées. De même, la base est alimentée avec la transcription de chaque ligne des carnets. Enfin, le lac de données du projet DataLAC comporte l'ensemble du corpus, les scripts Python d'extraction des métadonnées depuis les fichiers image pour l'alimentation de la base de métadonnées, le thésaurus ainsi que le système de gestion des métadonnées. Il est déployé sur les serveurs de l'infrastructure de recherche Huma-Num.

5 Conclusion

Le projet DataLAC est une preuve de concept de l'apport d'un lac de données et des méthodes d'intelligence artificielle pour l'exploitation d'archives archéologiques. Des améliorations ou des perspectives immédiates sont d'ores et déjà listées : tester d'autres modèles de segmentation et de transcription des pages plus récents, enrichir le thésaurus par de nouveaux concepts issus de l'analyse textuelle de l'ensemble des carnets, construire une carte géolocalisant les artefacts décrits dans les carnets, intégrer dans le lac les données de fouilles nativement numériques, etc. Le processus et l'architecture proposés dans le projet sont suffisamment généralisables pour qu'ils puissent être utilisés par d'autres sites de fouilles archéologiques. Le

8. https://thesaurus.mom.fr/?idt=th62

projet DataLAC trouvera ainsi une application directe : proposer une nouvelle voie d'organisation, de partage et d'analyse des données et d'archives en histoire et archéologie.

Nous remercions chaleureusement les étudiants du master humanités numériques de Lyon et Marianne Reboul, maîtresse de conférences à l'ENS de Lyon, qui les a supervisés dans le cadre d'un projet tutoré. Leur travail a été précieux pour l'avancement du projet DataLAC.

Références

Chagué, A., T. Clérice, J. Norindr, M. Humeau, B. Davoury, E. V. Kote, A. Mazoue, M. Faure, et S. Doat (2023). Manu McFrench, from zero to hero : impact of using a generic handwriting recognition model for smaller datasets. In *Digital Humanities 2023 : Collaboration as Opportunity*, Graz, Austria. Alliance of Digital Humanities Organizations and University of Graz.

Dixon, J. (2010). Pentaho, hadoop, and data lakes. Consulté le 19 octobre 2025. English

Gabay, S., A. Pinche, K. Christensen, et J.-B. Camps (2024). Segmonto : A controlled vocabulary to describe and process digital facsimiles. *Journal of Data Mining & Digital Humanities*, doi: 10.46298/jdmdh.12689.

Hai, R., S. Geisler, et C. Quix (2016). An intelligent data lake system. In *International Conference on Management of Data (SIGMOD 2016)*, San Francisco, CA, USA, pp. 2097–2100. ACM Digital Library.

Kiessling, B. (2020). A Modular Region and Text Line Layout Analysis System. In *2020 17th International Conference on Frontiers in Handwriting Recognition (ICFHR)*, Dortmund, Germany, pp. 313–318. IEEE, doi: 10.1109/ICFHR2020.2020.00064.

Sawadogo, P. N., E. Scholly, C. Favre, E. Ferey, S. Loudcher, et J. Darmont (2019). Metadata systems for data lakes : models and features. In *International Workshop on BI and Big Data Applications (BBIGAP@ADBIS 2019)*, Bled, Slovenia, pp. 440–451. Springer.

Scholly, E., P. N. Sawadogo, P. Liu, J. A. Espinosa-Oviedo, C. Favre, S. Loudcher, J. Darmont, et C. Nous (2021). Coining goldmedal : A new contribution to data lake generic metadata modeling. In *International Workshop on Design, Optimization, Languages and Analytical Processing of Big Data (DOLAP@EDBT 2021)*, Nicosia, Cyprus, pp. 31–40.

Summary

The DataLAC project focus on the use of artificial intelligence for the alignment, annotation and interpretation of heterogeneous documents enriched with semantic metadata and aggregated within a data lake. This project seeks to digitize, unify and make accessible over thirty years of field notes (1947–1977), scientific publications, and archives related to the Iberian archaeological site of Ullastret.These diverse materials are integrated into an interoperable data lake designed to manage the heterogeneity of sources and to support complex research queries. The DataLAC project thus constitutes a generalizable proof of concept, demonstrating the potential of data lake architectures and AI-driven methodologies for the exploitation and interpretation of archaeological archives.