

Structuration, exploration et valorisation d'archives archéologiques par l'intelligence artificielle au sein d'un lac de données

Rajae El-Idrissi¹ , Josefina Simon Reig² , Laura Romero² , Juba Agoun¹ ,
Jean-Pierre Girard³ , Gabriel de Prado² , Jérôme Darmont¹ , and
Sabine Loudcher¹ 

¹ Université Lumière Lyon 2, ERIC, Lyon, France

² Musée d'Archéologie de Catalogne, Site d'Ullastret, Espagne

³ Université Lumière Lyon 2, Archéorient, Lyon, France

Abstract

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.

Mots-clefs: Archéologie, Science des données, Lac de données, Métadonnées, Carnets de fouille

Keywords: Archaeology, Data science, Data lake, Metadata, Excavation diaries

1 Introduction

Alors que les premiers archéologues consignaient leurs observations principalement dans des carnets manuscrits, les protocoles actuels reposent sur des formulaires et des systèmes numériques de gestion de données. Toutefois, il n'existe pas encore de standard pour ces données : aucun consensus ne s'est imposé concernant les systèmes d'information archéologique, les outils utilisés (tableurs, bases FileMaker, bases relationnelles propriétaires ou open source, etc.) ni les pratiques de modélisation des données, malgré des méthodes de terrain relativement homogènes. La recherche archéologique repose également sur la comparaison d'artefacts présentant des caractéristiques similaires, à partir de photographies, d'illustrations ou de croquis issus de carnets de terrain. L'automatisation de l'analyse de ces corpus constitue un défi majeur, nécessitant des techniques de reconnaissance graphique appliquées à des archives scientifiques très hétérogènes (carnets numérisés, inventaires, dessins, cartes, plans). Idéalement, il faudrait aussi pouvoir interconnecter l'ensemble des données archéologiques (qu'elles soient manuscrites ou nativement numériques) afin de disposer d'une vision complète et cohérente du processus de fouille archéologique.

Le projet DataLAC vise à relever une partie de ces défis sur un site ibérique de la première moitié du VI^e siècle av. J.C. , le site d'Ullastret en Catalogne. Ce site est fouillé de façon continue depuis 1947 mais les documents des trente premières années de fouille

ne sont pas numériques et ils sont très hétérogènes (cahiers de fouille, archives photographiques, inventaires des objets découverts). Ils 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 faciliter le partage et l'analyse des données, le projet DataLAC mobilise le concept de lac de données qui vise à conserver ces dernières 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.

En d'autres termes, le projet DataLAC vise à développer une preuve de concept pour exploiter et analyser des corpus archéologiques en mobilisant des outils de science des données au sein d'un lac de données intégrant des sources hétérogènes et métadonnées.

Pour atteindre son objectif le projet DataLAC comprend quatre phases :

1. La numérisation des treize cahiers de fouilles manuscrits et de toute la documentation scientifique (rapports, photos, diapositives, plans, etc.). L'ensemble des documents numérisés constitue le corpus du projet.
2. La transcription automatique des 4000 pages des treize carnets grâce à la reconnaissance automatique d'écriture manuscrite à l'aide de modèles d'intelligence artificielle.
3. La modélisation et l'enrichissement des métadonnées décrivant les documents numérisés. Chaque carnet, page, ligne, 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 accessibles via un système de gestion des métadonnées composé d'une base de données relationnelle, d'une API 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 rassemble des archives qui retracent trois décennies de fouilles archéologiques du village ibérique de Puig de Sant Andreu à Ullastret¹. Au cœur 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 supplémentaires, 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 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.

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

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* »). Les archéologues impliqués dans le projet DataLAC ont défini un ensemble de descripteurs terminologiques fondés sur le vocabulaire utilisé dans les carnets de fouille, descripteurs qui permettent d’annoter les pages des carnets. Ils constituent des métadonnées riches sémantiquement et sont organisés dans un thésaurus² bilingue (catalan-castillan, qui sera ultérieurement apparié avec des concepts en français). Ce thésaurus constitue un outil 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 des différents de concepts identiques ou proches. À 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 (figure 1). Ce processus de reconnaissance de l’écriture manuscrite (ou HTR pour *Handwritten Text Recognition*) s’appuie sur des modèles d’apprentissage automatique affinés.

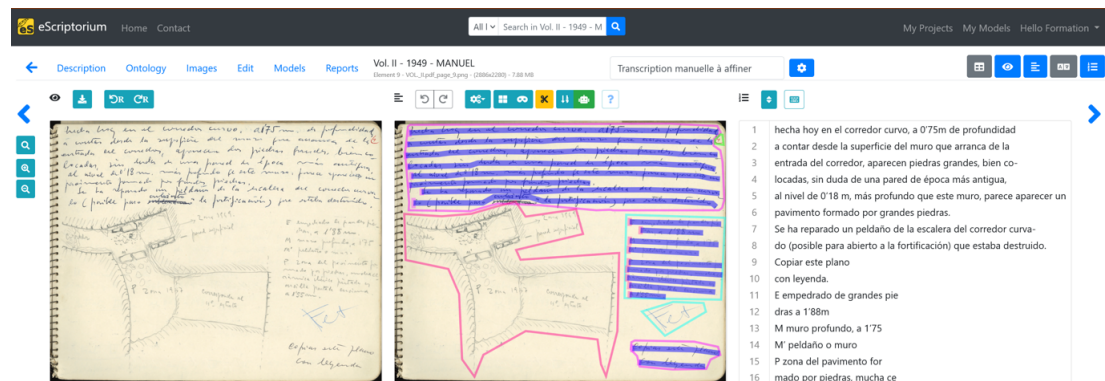


Figure 1 – Segmentation et transcription des pages des carnets de fouille

Toutes les opérations de segmentation et de transcription sont effectuées à l’aide de l’application eScriptorium³ et du moteur d’HTR Kraken, installés localement. Deux modèles pré-entraînés ont servi de point de départ avec un spécialisé dans la segmentation des zones de texte et des lignes, et l’autre pour la transcription proprement dite des zones segmentées. Les résultats sont exportés au format XML ALTO, avec des balises conformes aux principes de l’ontologie SegmOnto⁴ [7].

2. <https://thesaurus.mom.fr/?idt=th62>

3. <https://escriptorium.rich.ru.nl/> et <https://gitlab.com/scripta/escriptorium/>

4. <https://segmonto.github.io/>

3.1 Segmentation des pages

Le modèle de segmentation `b11a.mlmodel` est utilisé pour identifier automatiquement les lignes de texte dans les pages des carnets [2]. Ce modèle⁵, à base de réseaux de neurones convolutifs (CNN), est le modèle de segmentation de **Kraken** pré-entraîné à l’aide de toutes les données de Gallicorpora⁶.

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 (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 `b11a.mlmodel`. L’entraînement, réalisé sur une GPU, a nécessité environ 7 heures de calcul. Estimée par validation croisée, la proportion de pixels correctement classés est de 97,2%.

Le score moyen d’intersection (*Intersection over Union* IoU) vaut 36,6%, une valeur relativement faible, expliquée par la formation de contours flous autour des zones. En effet, les carnets 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. Néanmoins, l’IoU pondéré atteint 79%, indiquant une bonne performance globale du modèle notamment pour la zone principale contenant les lignes de texte.

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 à toutes les zones de texte segmentées [1]. 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,

5. <https://zenodo.org/records/14602569>

6. <https://github.com/Gallicorpora/Segmentation-and-HTR-Models>

7. <https://zenodo.org/record/6657809>

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 des 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

Pour lever les verrous informatiques liés au stockage, à l'interrogation, à l'analyse et à la visualisation des données hétérogènes, le projet utilise le concept de lac de données qui propose de stocker les données dans leur format d'origine, sans schéma prédéfini et sans transformation préalable à leur stockage [4 ; 6]. 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 le projet, nous utilisons la définition du concept de lac de données ainsi que les fonctionnalités, qu'un système de gestion de métadonnées doit avoir pour être complet, que nous avons proposées dans des travaux antérieurs [3 ; 5].

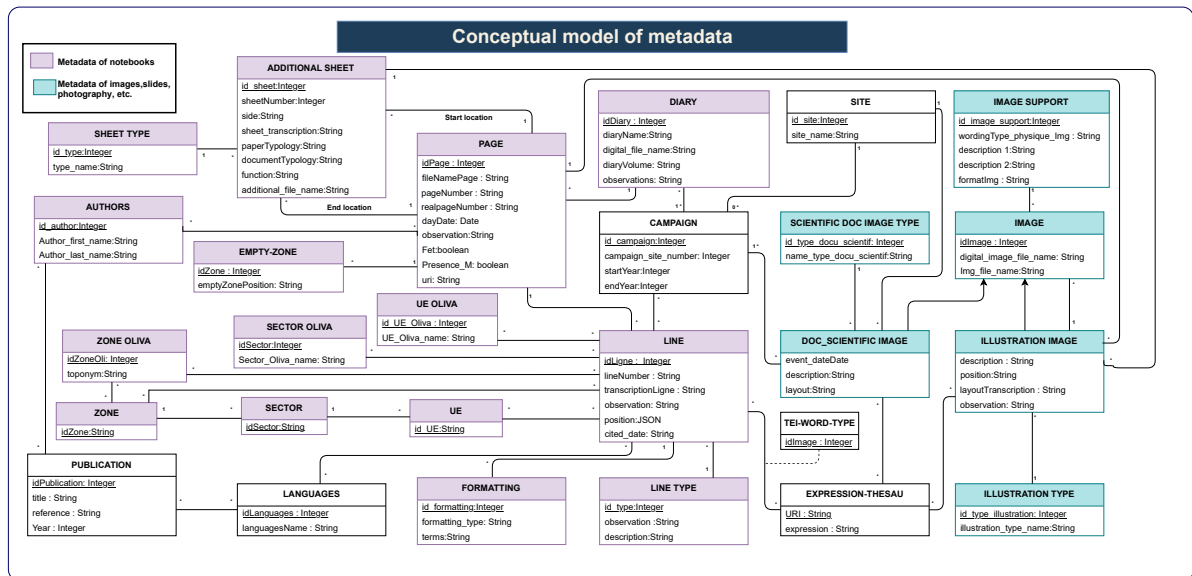


Figure 2 – Modèle conceptuel des métadonnées

La figure 2 représente le modèle conceptuel des métadonnées décrivant tous les documents numérisés à savoir les carnets, leurs pages, les illustrations et feuilles supplémentaires

présentes dans les carnets ainsi que toute la documentation scientifique. Chaque type de document correspondant à une classe dans le modèle conceptuel avec les attributs appropriés. La transcription automatique de chaque ligne de texte des pages est dans la classe ligne. Les mots-clés, désignés par *expression-thesau* dans la figure 2, sont piochés dans le thésaurus et décrivent sémantiquement tous les documents. Les zones, secteurs et unités stratigraphiques d’Oliva permettant de localiser les descriptions faites dans les lignes des pages sont également dans le thésaurus.

Ainsi toutes les documents du corpus sont décrits par une série de métadonnées et le thésaurus permet de les relier entre elles.

4.2 Architecture du lac de données

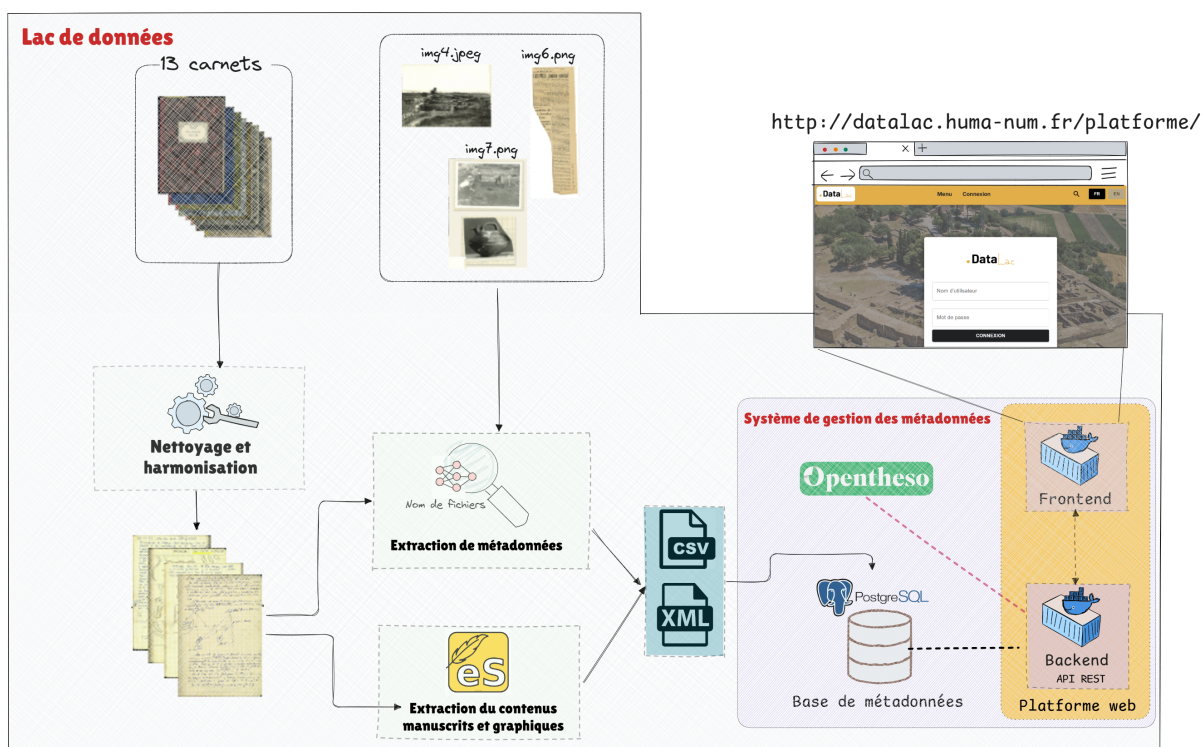


Figure 3 – Architecture du lac de données

Le système de gestion des métadonnées comprend une interface Web conçue pour la saisie/modification, la consultation et la mise en relation des métadonnées des documents du corpus (figure 3). L’interface communique avec le back-end via une API RESTful. Les métadonnées sont stockées dans une base de données relationnelle PostgreSQL. Le thésaurus, hébergé sur la plateforme Openthese⁸, est intégré dans le système de gestion de métadonnées. 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, tâche fastidieuse et chronophage, des scripts Python extraient des noms de fichiers les métadonnées de base

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

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 pages.

Dans le lac de données sont stockés l'ensemble des documents numérisés du corpus, mais aussi les fichiers résultant de la transcription automatique (un par page), les scripts, le thésaurus, le système de gestion des métadonnées (dont la base de métadonnées). Le lac est déployé sur les serveurs de l'IR* 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 envisagées : 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. 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 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.

Références

- [1] A., Chagué, T., Clérice, J., Norindr, M., Humeau, B., Davoury, E., Kote, A., Mazoue, M., Faure et S., Doat. « Manu McFrench, from zero to hero: impact of using a generic handwriting recognition model for smaller datasets ». In : *Digital Humanities 2023: Collaboration as Opportunity*. Alliance of Digital Humanities Organizations and University of Graz. Graz, Austria, juill. 2023. URL : <https://inria.hal.science/hal-04094241>.
- [2] B., Kiessling. « A Modular Region and Text Line Layout Analysis System ». In : *2020 17th International Conference on Frontiers in Handwriting Recognition (ICFHR)*. Dortmund, Germany : IEEE, sept. 2020, p. 313-318. DOI : 10.1109/ICFHR2020.2020.00064. URL : <https://hal.science/hal-04442992>.
- [3] E., Scholly, N., Sawadogo P., P., Liu, A., Espinosa-Oviedo J., C., Favre, S., Loudcher, J., Darmont et C., Nous. « 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, 2021, p. 31-40.
- [4] J., Dixon. « Pentaho, Hadoop, and Data Lakes ». Consulté le 19 octobre 2025. 2010. URL : <https://jamesdixon.wordpress.com/2010/10/14/pentaho-hadoop-and-data-lakes/>.
- [5] N., Sawadogo P., E., Scholly, C., Favre, E., Ferey, S., Loudcher et J., Darmont. « Metadata systems for data lakes: models and features ». In : *International Workshop on BI and Big Data Applications (BBIGAP@ADBIS 2019)*. Bled, Slovenia : Springer, 2019, p. 440-451.
- [6] R., Hai, S., Geisler et C., Quix. « An Intelligent Data Lake System ». In : *International Conference on Management of Data (SIGMOD 2016)*. San Francisco, CA, USA : ACM Digital Library, 2016, p. 2097-2100.

- [7] S., Gabay, A., Pinche, K., Christensen et JB., Camps. « SegmOnto: A Controlled Vocabulary to Describe and Process Digital Facsimiles ». English. In : *Journal of Data Mining & Digital Humanities* (déc. 2024). ISSN : 2416-5999. DOI : 10.46298/jdmdh.12689. URL : <https://jdmdh.episciences.org/12689>.