DMP du projet "DataCatalogue"

Plan de gestion de données créé à l'aide de DMP OPIDoR, basé sur le modèle "ANR - Modèle de PGD (français)" fourni par Agence nationale de la recherche (ANR).

Renseignements sur le plan

Titre du plan	DMP du projet "DataCatalogue"
Version	Version initiale
Langue	fra
Date de création	2021-11-04
Date de dernière modification	2021-11-05

Renseignements sur le projet

Titre du projet

DataCatalogue

Résumé

La transition entre une numérisation patrimoniale des collections de catalogues et l'inscription dans la dynamique "collection as data" est fortement attendue par la communauté scientifique. Pour qu'elle soit rendue possible, le développement d'outils d'extraction de la structure logique des catalogues de vente et leur mise à disposition auprès des institutions et des équipes de recherche est un prérequis.

L'objectif est :

- Passer d'une numérisation en mode image à une base de données textuelles et requêtable.
- Entrainer GROBID pour la segmentation permettant de "zoner" les catalogues de ventes et pour chacune de ses zones, leurs attribuer une fonction "n° des lots, description du lot".
- Un point d'attention serait le cas des images dans les catalogues de ventes. Il s'agira de relier les photographies de lot avec le numéro de lot de la vente correspondant (l'image est très souvent quelques pages après la description du lot, le lien se faisant sur le n° et la légende de l'image).
- Mettre à disposition des chercheurs une interface de recherche minimale permettant de requêter sur les zones segmentées.

Sources de financement

- : Convention cadre Inria-Ministère de la Culture. Co-financement Ministère de la culture, Bibliothèque nationale de France et Institution National d'Histoire de l'Art.

Date de début

2021-10-01

Date de fin

2022-09-30

Partenaires

- Inria (196718247G)
- Institut National d'Histoire de l'Art (200120913G)
- Bibliothèque nationale de France ()

Produits de recherche :

1. Corpus d'entraînement de fichiers XML TEI pour le module GROBID catalogues de vente (ODD)
2. Modèle d'analyse structurée pour les catalogues de vente GROBID (Modèle)
3. Corpus de fichiers XML TEI encodés suite à l'entraînement du modèle GROBID (Jeu de données)
4. Interface de requêtage avec TEI Publisher (Service)
5. ODD catalogues de vente (ODD)

Contributeurs

Nom	Affiliation	Rôles
Scheithauer Hugo	Inria	• Responsable du plan de gestion de données
Romary Laurent	Inria	• Coordinateur du projet • Personne contact pour les données (Modèle catalogue, Corpus entraînement, Corpus encodé, ODD catalogue, Interface requêtage)

DMP du projet "DataCatalogue"

1. Description des données et collecte ou réutilisation de données existantes

ODD catalogues de vente

Un modèle de données de type ODD, issu d'une customisation de la TEI, sera produit pour encoder les catalogues de vente.

L'ODD est un fichier au format XML.

Corpus d'entraînement de fichiers XML TEI pour le module GROBID catalogues de vente

Les numérisations des catalogues de vente mises à disposition par la Bibliothèque nationale de France et l'Institut national d'histoire de l'art, avec leurs textes obtenus grâce à des outils d'OCR, seront utilisés pour produire des encodages XML TEI. Ils serviront à constituer un corpus d'entraînement pour un modèle GROBID.

Le corpus d'entraînement rassemblera des fichiers XML TEI.

Modèle d'analyse structurée pour les catalogues de vente GROBID

Un modèle d'analyse structurée GROBID des catalogues de vente sera produit à partir du corpus d'entraînement.

Le modèle sera implémenté dans un module GROBID, écrit en Java.

Corpus de fichiers XML TEI encodés suite à l'entraînement du modèle GROBID

Les numérisations des catalogues de vente mises à disposition par la Bibliothèque nationale de France et l'Institut national d'histoire de l'art, avec leurs textes obtenus grâce à des outils d'OCR, seront utilisés pour produire des encodages XML TEI grâce à l'outil d'analyse structurée GROBID et du modèle entraîné pour cette tâche dans le cadre du projet.

Au total, 9100 catalogues de vente sont aujourd'hui numérisés. Il y aura autant de fichiers XML TEI produits grâce à l'utilisation de ce modèle, que de numérisations disponibles.

Interface de requêtage avec TEI Publisher

Une interface de requêtage sera produite à partir d'un paramétrage du logiciel open-source TEI Publisher.

Des fichiers HTML, CSS et une ODD seront produits pour paramétrer TEI Publisher.

2. Documentation et qualité des données

ODD catalogues de vente

Le standard de métadonnées utilisé sera la TEI.

L'ODD produite fera office de schéma de contrôle pour les fichiers XML TEI produits dans le cadre du projet.

Corpus d'entraînement de fichiers XML TEI pour le module GROBID catalogues de vente

Le standard de métadonnées utilisé sera la TEI.

Les données XML seront contrôlées par validation avec le fichier ODD issu de la customisation TEI spécifique au projet.

Modèle d'analyse structurée pour les catalogues de vente GROBID

Le modèle produit sera documenté sur l'organisation Github du projet.

N/A

Corpus de fichiers XML TEI encodés suite à l'entraînement du modèle GROBID

Le standard de métadonnées utilisé sera la TEI.

Les données XML seront contrôlées par validation du schéma de la customisation du schéma TEI spécifique au projet.

Interface de requêtage avec TEI Publisher

Le paramétrage de TEI Publisher s'appuiera sur la documentation déjà existante et spécifique au logiciel.

La base de données eXist-db servant à faire fonctionner l'application sera utilisée comme outil de contrôle.

3. Stockage et sauvegarde pendant le processus de recherche

Les données et métadonnées seront stockées et sauvegardées au fur et à mesure des expérimentations et avancées dans une organisation Github au nom du projet (https://github.com/DataCatalogue/). Un répertoire Gitlab sur l'instance d'Inria sera également branché en tant que "remote" au répertoire Github principal, afin d'assurer l'archivage du projet.
Un espace Sharedocs (Huma-Num) a également été mis en place afin de stocker les fichiers et permettre la collaboration entre les acteurs du projet.

Tous les acteurs du projet auront accès aux données via Sharedocs. L'accès à cet espace de stockage leur est réservé.
Tous les acteurs du projet pourront contribuer à l'organisation Github, qui assure un système d'intégration continue permettant de versionner les données, et d'en garder trace sur le temps long. L'accès à ce répertoire est public.

---

4. Exigences légales et éthiques, codes de conduite

Il n'y aura pas de données à caractère personnel qui seront traitées.

Les répertoires de l'organisation Github DataCatalogue seront placés sous licence Creative Commons Attribution 4.0 International et respecteront la charte Heritage Data Reuse (https://datacharter.hypotheses.org/mission-statement).

Il n'y a pas de comité éthique mis en place pour le projet.

---

5. Partage des données et conservation à long terme

Les données seront partagées publiquement, sous licence open-source et selon les principes de la charte Heritage Data Reuse (https://datacharter.hypotheses.org/mission-statement). Les données seront rendues publiques au moment où elles sont publiées sur Github, donc au cours de l'avancée du projet.

Ne seront conservées dans l'organisation Github que les données utiles à l'avancée du projet. L'organisation se situe à l'adresse suivante : https://github.com/DataCatalogue

---

Les données sont accessibles avec un accès internet et un navigateur, afin de se rendre sur le répertoire Github.

Question sans réponse.

---

6. Responsabilités et ressources en matière de gestion des données

Les responsables de la gestion des données sont :

- M. Laurent Romary, directeur de recherche, Inria - ALMAnaCH
- M. Hugo Scheithauer, ingénieur recherche et développement, Inria - ALMAnaCH

Une des missions de l'ingénieur recruté pour le projet consiste à rendre les données FAIR au cours du projet.