Le projet DataCatalogue

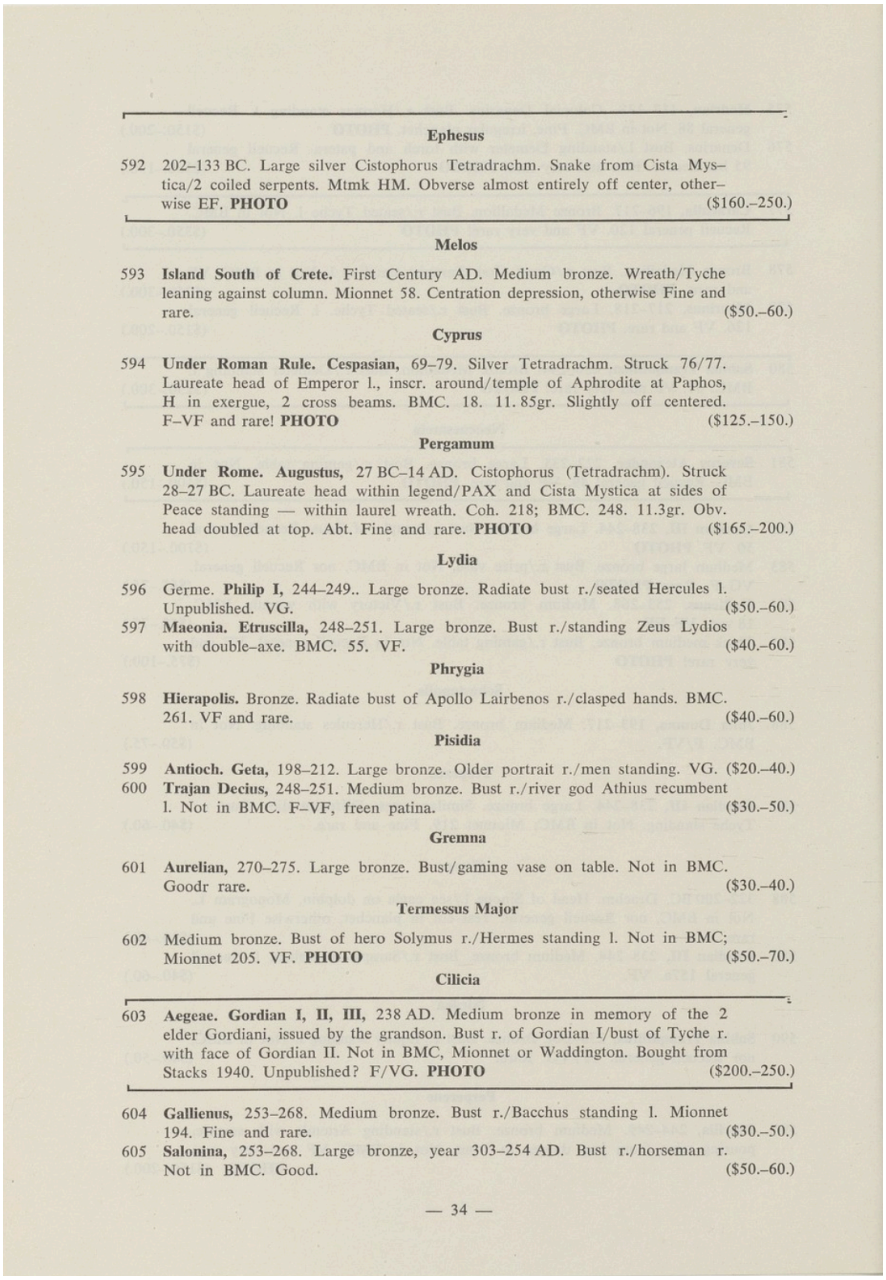
Les catalogues de ventes publiques d'œuvres d'art ou d'objets archéologiques représentent une ressource particulièrement abondante de données. Si le marché passe aujourd'hui en grande partie par internet, les catalogues de vente imprimés restent une ressource exceptionnellement riche dans les collections de la BnF comme de l'INHA, fondamentale pour beaucoup de recherches historiques. À la BnF, la collecte a permis de rassembler des catalogues dont le plus ancien date de 1726 tandis que les entrées se poursuivent chaque année par le biais du Dépôt légal aussi bien que par d'autres modes d'acquisition. Les fascicules d'ores et déjà en ligne sont divers tant par leur date d'édition (depuis 1730 jusqu'à aujourd'hui) que par leur langue (français, anglais, italien, allemand, etc., avec certains catalogues multilingues). Le fonds de l'INHA, qui comprend les catalogues de la bibliothèque d'art et d'archéologie Jacques Doucet et ceux provenant de la bibliothèque centrale des musées nationaux, regroupe plus de 160 000 catalogues de vente, dont les plus anciens datent du 17e siècle. La collection s'enrichit d'environ 1 000 documents par an.

Ces collections sont très demandées par les chercheurs en histoire, histoire de l'art et archéologie mais aussi par certaines catégories de professionnels comme les marchands d'art. Elles représentent une source documentaire précieuse par le nombre d'œuvres qu'elles font connaître dont certaines ont ensuite disparu et ne sont plus connues que par le catalogue de vente. Pour beaucoup d'autres objets, la vente n'est qu'une étape entre deux collections, qu'elles soient publiques ou privées. De ce fait, les catalogues de vente sont une source essentielle de l'histoire des collections et des collectionneurs. Plus simplement, les catalogues de vente sont une ressource iconographique particulièrement riche. Concernant les objets sériels (monnaies, médailles, objets archéologiques comme les poids de marchés antiques, les jetons, etc.), les catalogues de vente fournissent une part très substantielle des corpus constitués par le seul nombre des objets qu'ils documentent.

Dans le contexte de cette activité scientifique, les collections de la BnF et de l'INHA sont centrales par leur nombre mais aussi par les annotations portées sur de nombreux catalogues :

prix de vente, nom de l'acheteur, etc. Cependant, de nombreuses limites sont imposées aux usages possibles de ces collections numérisées :

- Du fait du processus historique de numérisation opéré au sein des bibliothèques, les données numériques relatives à ces collections sont réparties entre métadonnées conservées dans des catalogues et images et textes diffusés via des portails web de bibliothèques numériques.
- Quand il est disponible, le texte obtenu par le moyen d'une océrisation qui a pour échelle la page numérisée, et ne permet pas d'accéder aux grains d'information utiles (ventes, objets), ni a fortiori aux données atomiques (prix, nom, date, etc.).
- Le silotage des collections entre les différents détenteurs est également un facteur limitant certaines types d'études, notamment les approches quantitatives.



Ephesus

592 202–133 BC. Large silver Cistophorus Tetradrachm. Snake from Cista Mystica/2 coiled serpents. Mtmk HM. Obverse almost entirely off center, otherwise EF. **PHOTO** ($160.–250.)

Melos

593 **Island South of Crete.** First Century AD. Medium bronze. Wreath/Tyche leaning against column. Mionnet 58. Centration depression, otherwise Fine and rare. ($50.–60.)

Cyprus

594 **Under Roman Rule. Cespasian,** 69–79. Silver Tetradrachm. Struck 76/77. Laureate head of Emperor l., inscr. around/temple of Aphrodite at Paphos, H in exergue, 2 cross beams. BMC. 18. 11.85gr. Slightly off centered. F–VF and rare! **PHOTO** ($125.–150.)

Pergamum

595 **Under Rome. Augustus,** 27 BC–14 AD. Cistophorus (Tetradrachm). Struck 28–27 BC. Laureate head within legend/PAX and Cista Mystica at sides of Peace standing — within laurel wreath. Coh. 218; BMC. 248. 11.3gr. Obv. head doubled at top. Abt. Fine and rare. **PHOTO** ($165.–200.)

Lydia

596 Germe. **Philip I,** 244–249.. Large bronze. Radiate bust r./seated Hercules l. Unpublished. VG. ($50.–60.)

597 **Maeonia. Etruscilla,** 248–251. Large bronze. Bust r./standing Zeus Lydios with double-axe. BMC. 55. VF. ($40.–60.)

Phrygia

598 **Hierapolis.** Bronze. Radiate bust of Apollo Lairbenos r./clasped hands. BMC. 261. VF and rare. ($40.–60.)

Pisidia

599 **Antioch. Geta,** 198–212. Large bronze. Older portrait r./men standing. VG. ($20.–40.)

600 **Trajan Decius,** 248–251. Medium bronze. Bust r./river god Athius recumbent l. Not in BMC. F–VF, freen patina. ($30.–50.)

Gremna

601 **Aurelian,** 270–275. Large bronze. Bust/gaming vase on table. Not in BMC. Goodr rare. ($30.–40.)

Termessus Major

602 Medium bronze. Bust of hero Solymus r./Hermes standing l. Not in BMC; Mionnet 205. VF. **PHOTO** ($50.–70.)

Cilicia

603 **Aegeae. Gordian I, II, III,** 238 AD. Medium bronze in memory of the 2 elder Gordiani, issued by the grandson. Bust r. of Gordian I/bust of Tyche r. with face of Gordian II. Not in BMC, Mionnet or Waddington. Bought from Stacks 1940. Unpublished? F/VG. **PHOTO** ($200.–250.)

604 **Gallienus,** 253–268. Medium bronze. Bust r./Bacchus standing l. Mionnet 194. Fine and rare. ($30.–50.)

605 **Salonina,** 253–268. Large bronze, year 303–254 AD. Bust r./horseman r. Not in BMC. Good. ($50.–60.)

— 34 —

https://gallica.bnf.fr/ark:/12148/bd6t54182207f

Catalogue de vente, Consignments from the (Victor H.) Weill collection. United States coins. The Vatican duplicates of ancient coins, Hans M. F. Schulman, 1972.

Le projet DataCatalogue tente de répondre à ces questions. Il a été mis en place dans le cadre d'un partenariat entre l'Inria[1], la Bibliothèque nationale de France[2] (BnF) et l'Institut national

[1] https://inria.fr/fr
[2] https://www.bnf.fr/fr

d'histoire de l'art[3] (INHA), conclu dans le cadre de la Convention Culture-Inria[4] (ministère de la Culture et Inria). L'objectif du projet est d'automatiser la transformation des catalogues de ventes numérisés et mis à disposition dans les bibliothèques numériques de la BnF et de l'INHA dans un format structuré permettant une recherche plus fine basée sur le contenu sémantique textuel de ces documents, au-delà du niveau de l'image.

La première phase du projet (12 mois, 2021-2022) a permis d'établir un modèle de données pour représenter les catalogues de vente et leurs différents niveaux d'information, qui a ensuite été approfondi et amélioré au cours de deuxième phase (12 mois, 2023-2024). Celui-ci a été créé à partir d'une personnalisation du standard de données XML-TEI (*Text Encoding Initiative*), et vise à pouvoir modéliser les catalogues selon un niveau de granularité aussi fin que possible. Il se veut également suffisamment générique pour pouvoir englober toute la diversité des formats des catalogues à travers le temps et les domaines de vente. Les catalogues de vente, et plus largement presque tous types de documents, peuvent être segmentés à deux niveaux distincts : un premier qu'on nommera macro-segmentation, qui concerne la mise en page du document (titre, paragraphes, entrées de catalogues, notes de bas de page, etc.) ; et la micro-segmentation, qui traite du contenu sémantique contenu dans le texte, par exemple dans le cas des catalogues, les éléments constitutifs d'une entrée (nom de l'objet vendu, description, prix, commentaire de conservation, etc.).

Après une prise en main de GROBID, un logiciel destiné à structurer le contenu issus d'articles scientifiques, et de son architecture logicielle[5], le développement d'un module dédié à la segmentation des catalogues de vente a été entrepris. Le fonctionnement de GROBID est basé sur une cascade de modèles d'apprentissage machine qui vont à tour de rôle segmenter de plus en plus finement un document.

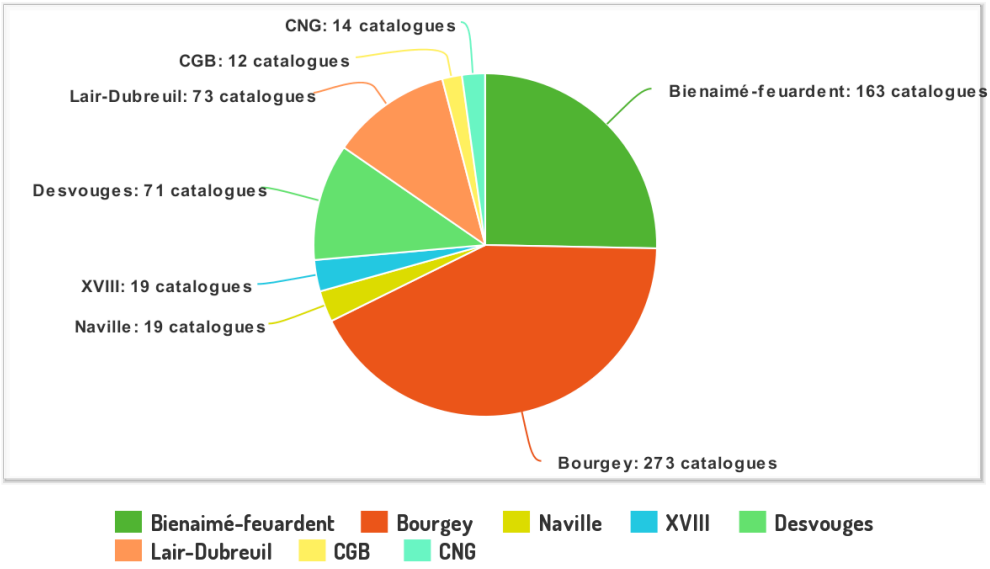
Chaque niveau de segmentation implique la création d'un corpus annoté et l'entraînement d'un modèle à partir de celui-ci. Pour faciliter ce processus, un corpus a été échantillonné à partir des collections de la BnF et de l'INHA. Il se veut être représentatif en termes de datation, de type de vente, de typographie et de typologie, et intègre différents niveaux de bruit d'OCR, afin

[3] https://www.inha.fr/

[4] https://www.culture.gouv.fr/fr/catalogue-des-demarches-et-subventions/subvention/convention-culture-inria.

[5] https://grobid.readthedocs.io/en/latest/Introduction/

d’envisager des modèles de segmentation aussi robuste que possible face à la diversité présente dans les collections de catalogues. Au total, 713 catalogues ont été échantillonnés.



Échantillon DataCatalogue

Un modèle de segmentation haut niveau, destiné à structurer un catalogue en quatre parties, à savoir les parties liminaires, le corps du texte contenant les entrées, la quatrième de couverture, et les annexes, a été entraîné. Pour cela, 644 catalogues contenus dans l’échantillon ont été annotés pour entraîner et tester la robustesse du modèle. Ses performances ont permis le développement du niveau de segmentation suivant, à savoir le découpage des niveaux de titre et des différentes notices contenues dans le corps des catalogues.

Cependant, les conclusions de la première phase ont montré que le bruit parfois généré par l’océrisation des catalogues, comme des erreurs d’OCR, des mots hallucinés par le modèle, ou des mots déconstruits, a affecté la performances des modèles entraînés et les a empêché d’atteindre des scores satisfaisants. La deuxième phase du projet DataCatalogue (2023-2024) a donc été conçue de sorte à pouvoir contourner cette limitation.

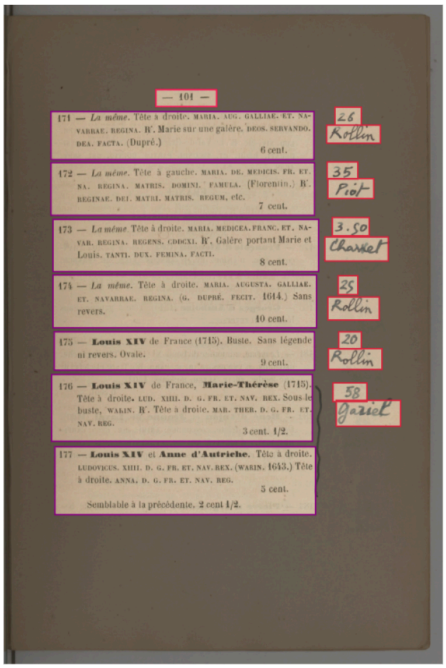
Suivant l'exemple d'expériences concluantes menées au sein du laboratoire ALMAnaCH à Inria dans le cadre du projet COLaF (Corpus et Outils pour les Langues de France)[6], cette deuxième phase s'est concentrée sur l'étape de macro-segmentation en s'appuyant cette fois-ci uniquement sur les numérisations, et non leurs océrisations. DataCatalogue est ainsi venu s'insérer au sein d'une campagne d'annotation plus grande, destinée à créer un large corpus diachronique de documents historiques dont la mise en page est annotée, et qui se nomme LADaS (*Layout Analysis Dataset with SegmOnto*[7]. Deux pages aléatoires ont été prélevées dans les 713 catalogues de l'échantillon puis annotées de sorte à créer un subset DataCatalogue de ce plus large corpus. Ce jeu de données a ensuite permis d'entraîner un modèle de détection d'objet nommé YOLO[8], qui se base uniquement sur des informations visuelles, au contraire de GROBID qui se sert d'informations textuelles et visuelles.

Les données ont été annotées en suivant les recommandations du vocabulaire contrôlé SegmOnto[9], destiné à homogénéiser la description de mise en page des manuscrits, et utilisé dans le cadre de la campagne LADaS. Le vocabulaire SegmOnto a été adapté pour les besoins du projet aux catalogues de vente. Un guide d'annotation a été conçu de sorte à identifier avec précision les éléments structurant des catalogues de vente, comme les notices par exemple[10]. L'utilisation de SegmOnto dans la création de données d'entraînement permet de rendre le subset DataCatalogue interopérable, notamment avec LADaS, mais aussi dans le futur avec d'autres jeux de données, et permet de respecter les principes FAIR (*Findable, Accessible, Interoperable, Reusable*)[11] de manière générale.

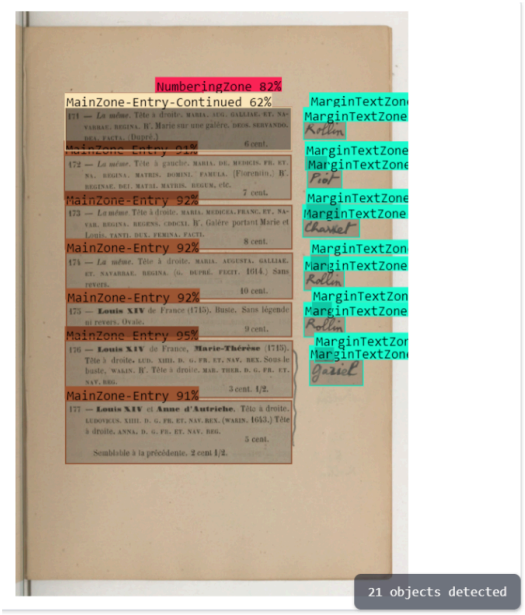
Les résultats de l'entraînement de ce modèle YOLO pour DataCatalogue se sont montrés satisfaisants, et permettent aujourd'hui une macro-structuration automatique pour les catalogues de vente.

---

[6] https://colaf.huma-num.fr/
[7] https://github.com/DEFI-COLaF/LADaS
[8] https://docs.ultralytics.com/fr
[9] https://segmonto.github.io/
[10] https://github.com/DataCatalogue/datacat-object-detection-dataset/blob/main/DataCat_AnnotationGuide.md
[11] https://www.go-fair.org/fair-principles/



Vérité de terrain

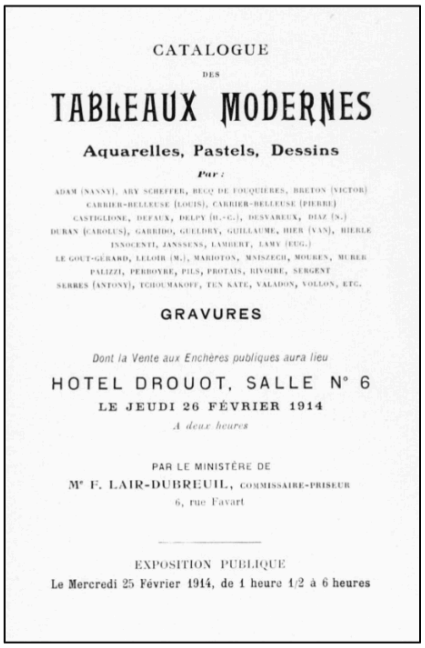


Prédiction

Comparaison entre la vérité de terrain et la prédiction du modèle

Enfin, le modèle de données TEI a été approfondi, et une ODD (*One Document Does it all)*[12], un fichier permettant le contrôle du modèle, a été créé pour l'encodage des catalogues de vente. La collaboration entre les compétences techniques de modélisation à Inria et l'expertise des catalogues de vente des agents de l'INHA et de la BnF ont permis d'identifier des éléments clés de ce type de document.

[12] https://github.com/DataCatalogue/datacat-tei/tree/main/ODD



CATALOGUE

DES

TABLEAUX MODERNES

Aquarelles, Pastels, Dessins

Par :

ADAM (NANNY), ARY SCHEFFER, BACQ DE FOUQUIÈRES, BRETON (VICTOR) CARRIER-BELLEUSE (LOUIS), CARRIER-BELLEUSE (PIERRE) CASTIGLIONE, DEFAUX, DELPY (H.-C.), DESVAREUX, DIAZ (N.) DURAN (CAROLUS), GARRIDO, GUELDRY, GUILLAUME, HIER (VAN), HIERLE INNOCENTI, JANSSENS, LAMBERT, LAMY (EUG.) LE GOUT-GÉRARD, LELOIR (M.), MARIOTON, MANISZECH, MOUREN, MURER PALIZZI, PERBOYRE, PILS, PROTAIS, RIVOIRE, SERGENT SERRES (ANTONY), TCHOUMAKOFF, TEN KATE, VALADON, VOLLON, ETC.

GRAVURES

Dont la Vente aux Enchères publiques aura lieu

HOTEL DROUOT, SALLE N° 6

LE JEUDI 26 FÉVRIER 1914

A deux heures

PAR LE MINISTÈRE DE

Me F. LAIR-DUBREUIL, COMMISSAIRE-PRISEUR

6, rue Favart

EXPOSITION PUBLIQUE

Le Mercredi 25 Février 1914, de 1 heure 1/2 à 6 heures

Lair-Dubreuil, 1914

```xml
<front>
	<titlePage>
		<docTitle>
			<titlePart type="main">CATALOGUE <lb/>DES <lb/> TABLEAUX MODERNES</titlePart>
			<titlePart type="sub">Aquarelles, Pastels, Dessins</titlePart>
			<titlePart type="artists">Par : <lb/><name>ADAM (NANNY)</name>, <name>ARY SCHEFFER</name>,
				<name>BACQ DE FOUQUIERES</name>, <name>BRETON (VICTOR)</name>
			<lb/><name>CARRIER-BELLEUSE (LOUIS)</name>, <name>CARRIER-BELLEUSE (PIERRE)</name>
			<lb/><name>CASTIGLIONE</name>, <name>DEFAUX</name>, <name>DELPY (H.-C.)</name>,
				<name>DESVAREUX</name>, <name>DIAZ (N.)</name>
			<lb/><name>DURAN (CAROLUS)</name>, <name>GARRIDO</name>, <name>GUELDRY</name>,
				<name>GUILLAUME</name>, <name>HIER (VAN)</name>, <name>HIERLE</name>
			<lb/><name>INNOCENTI</name>, <name>JANSSENS</name>, <name>LAMBERT</name>, <name>LAMY
				(EUG.)</name>
			<lb/><name>LE GOUT-GÉRARD</name>, <name>LELOIR (M.)</name>, <name>MAIOTON</name>,
				<name>MNISZECH</name>, <name>MOUREN</name>, <name>MURER</name>
			<lb/><name>PALIZZI</name>, <name>PERBOYRE</name>, <name>PILS</name>,
				<name>PROTAIS</name>, <name>RIVOIRE</name>, <name>SERGENT</name>
			<lb/><name>SERRES (ANTONY)</name>, <name>TCHOUMAKOFF</name>, <name>TEN KATE</name>,
				<name>VALADON</name>, <name>VOLLON</name>, ETC.</titlePart>
			<titlePart type="sub">GRAVURES</titlePart>
			<titlePart type="sales_place">Dont la Vente aux Enchères publiques aura lieu
				<lb/><placeName>HOTEL DROUOT</placeName>, SALLE N°6 <lb/><date>LE JEUDI 26 FÉVRIER
				1914</date><lb/><time>A deux heures</time></titlePart>
		</docTitle>

		<byline>PAR LE MINISTÈRE DE <lb/><name>Me F. LAIR-DUBREUIL</name>, COMMISSAIRE-PRISEUR <lb/>
			<address>
				<addrLine>6, rue Favart</addrLine>
			</address>
		</byline>

		<metamark style="line"/>
		<titlePart><desc>EXPOSITION PUBLIQUE</desc>
			<lb/><date>Le Mercredi 25 Février 1914</date>, <time>de 1 heure 1/2 à 6 heures</time>
		</titlePart>
	</titlePage>
</front>
```

ExampleFile_Lair-Dubreuil_CV02553_19140226_f3.xml

Exemple d'encodage TEI d'une page de titre de catalogue de ventes

Le travail réalisé dans le cadre du projet est disponible en libre accès et documenté sur GitHub[13]. Le jeu de données et le modèle YOLO "DataCatalogue" sont disponibles sur la plateforme d'annotation Roboflow[14].

Bibliographie et communications autour du projet

Sarah Bénière. DataCatalogue : Restructurer automatiquement les catalogues de ventes. M2 TNAH - Panorama de projets, Jan 2024, Paris, France. 2024. ⟨hal-04430891⟩

Sarah Bénière, Hugo Scheithauer, Juliette Janes, Laurent Romary. An ODD Schema for a Sustainable Encoding of Catalog Objects. TEI 2024 – Texts, Languages and Communities, Universidad del Salvador, Oct 2024, Buenos Aires, Argentina. ⟨hal-04754028⟩

[13] https://github.com/DataCatalogue
[14] https://app.roboflow.com/datacatalogue/macro-segmentation/overview

Thibault Clérice. You Actually Look Twice At it (YALTAi): using an object detection approach instead of region segmentation within the Kraken engine. 2023. ⟨hal-03723208v2⟩

Thibault Clérice, Juliette Janes, Hugo Scheithauer, Sarah Bénière, Laurent Romary, et al.. Layout Analysis Dataset with SegmOnto. DH2024 - Annual conference of the Alliance of Digital Humanities Organizations, ADHO, Aug 2024, Washington DC, United States. ⟨hal-04513725⟩

Ariane Pinche, Simon Gabay, Jean-Baptiste Camps. SegmOnto : Vocabulaire contrôlé pour décrire les manuscrits et les imprimés. Segmenter et annoter les images : déconstruire pour reconstruire, Nov 2022, Paris, France. ⟨hal-03930487⟩

Laurent Romary, Hugo Scheithauer. DataCatalogue : enjeux et réalisations. Un outil numérique pour interroger les catalogues de vente : le projet DataCatalogue, Oct 2022, Paris, France. ⟨hal-03829309⟩

Hugo Scheithauer, Sarah Bénière, Laurent Romary. Automatic retro-structuration of auction sales catalogs layout and content. DH2024 - Reinvention and Responsibility, Alliance of Digital Humanities Organizations, Aug 2024, Washinghton DC, United States. ⟨hal-04547239⟩