Enrichir le patrimoine écrit archivistique grâce aux technologies numériques : Ingénierie du projet LectAuRep (Lecture automatique de répertoires)

Aurélia Rostaing, Hugo Scheithauer

#dhnord2022
TRAVAILLER EN HUMANITÉS NUMÉRIQUES
collaborations, complémentarités et tensions

Enrichir le patrimoine écrit archivistique grâce aux technologies numériques

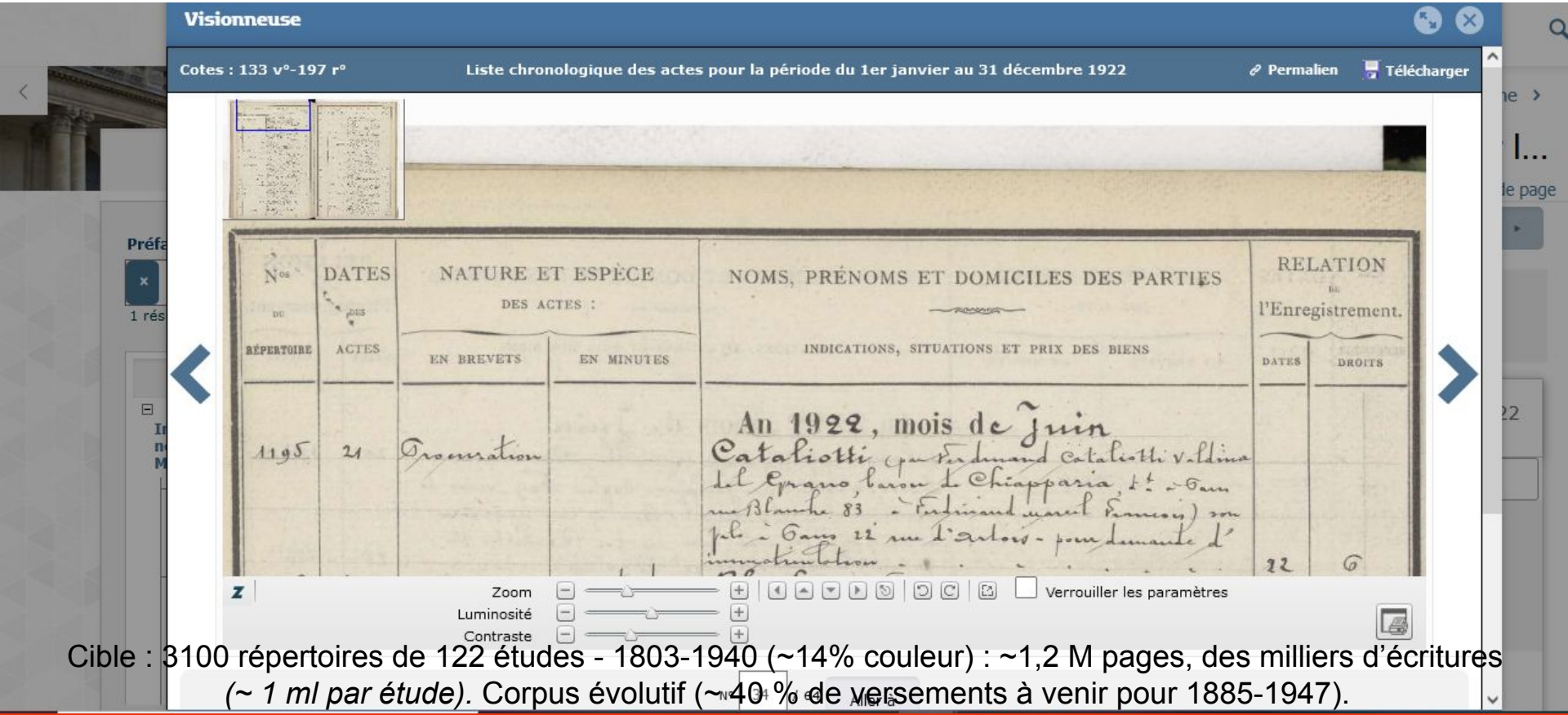
Ingénierie du projet LectAuRep

(Lecture automatique de répertoires)

Aurélia Rostaing (Archives nationales) - Hugo Scheithauer (ALMAnaCH, Inria)

21 juin 2022

A quoi sert un répertoire de notaire ?



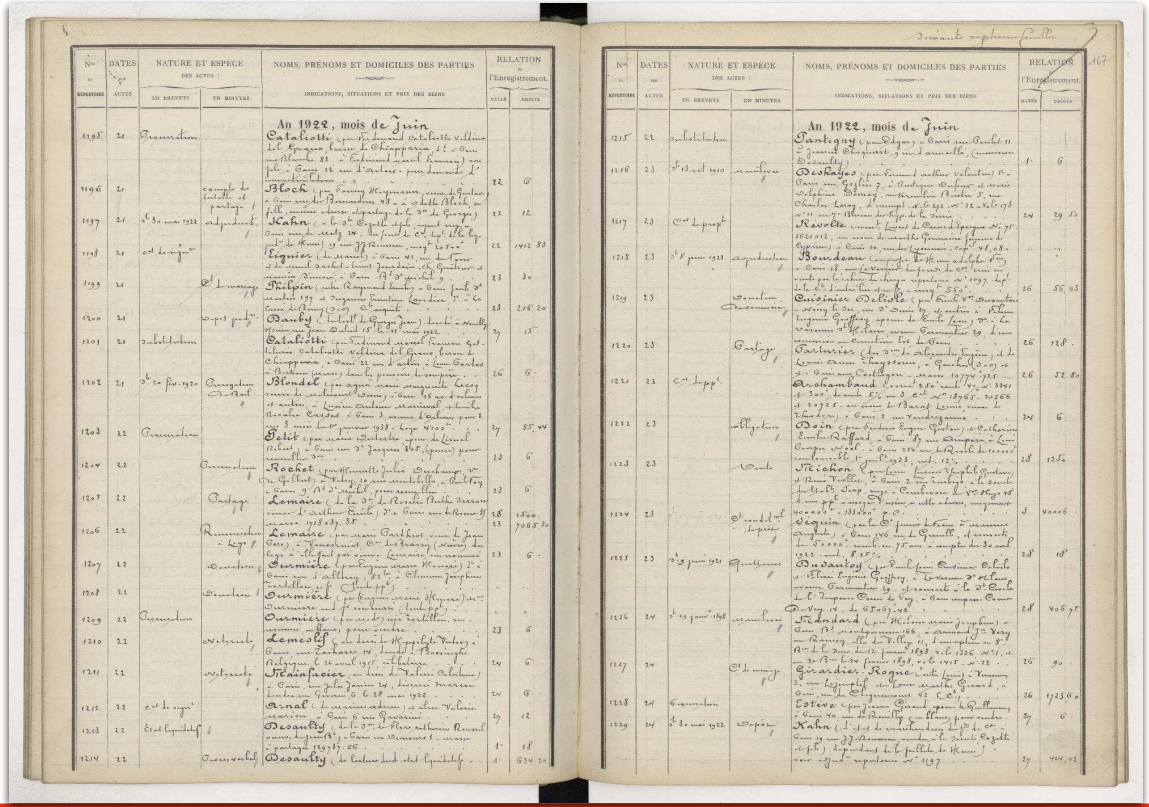
Cible : 3100 répertoires de 122 études - 1803-1940 (~14% couleur) : ~1,2 M pages, des milliers d'écritures *(~ 1 ml par étude)*. Corpus évolutif (~40 % de versements à venir pour 1885-1947).

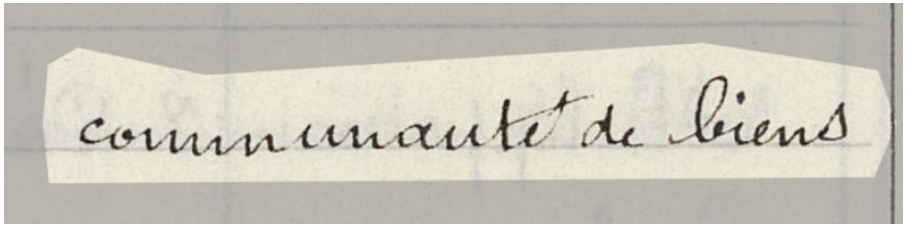
LectAuRep pourquoi ?

Fournir au public un service de recherche textuelle puis d'exploration visuelle de données
au sein d'images numériques de répertoires de notaires, dans le cadre d'une politique d'instruments de recherche.

Pas de répertoire
(~ 11,6 % des articles fin XVe-1885)
→ analyse pièce à pièce
des minutes originales

Répertoire
→ traitement par HTR
des images de répertoires
(+ TAL, KWS, NER...)





Un pot commun au service d’un réseau du patrimoine écrit (archives et GLAM), avec une prise en charge du patrimoine numérisé

> mutualisation de données, modèles et méthodes documentés, produits avec un logiciel libre

Vidéo MuseumWeek - 13/06/22

https://www.archives-nationales.culture.gouv.fr/innovation-et-collaboration-au-service-du-patrimoine#lectaurep

LectAuRep : résultats du projet

- **Échantillonnage de 4 lots** (autant de corpus)
  - ~ 1067 doubles pages transcrites (une centaine de mains), 286 relues (moins d'un tiers) (échantillon transcrit : 1,11 ‰ de la cible).

- **Données réutilisables**
  - 31 401 lignes de vérité terrain (de 239 images) disponibles sur HTR-United (RE, CM/SD, Bronod)

- **Modèles réutilisables (perfectibles)**
  - segmentation des lignes (1), segmentation des régions (1)
  - reconnaissance de texte (4) : CER inférieurs à 10, voire à 5 %
    - génériques (dizaines de mains) : "générique" (CER 9 %) et "random set" (CER 10 %)
    - spécifiques : "Bronod" (une main, CER 5 %) et "Contrats de Mariage" (~ 10 mains, CER 3 %)

- **Documentation** : conventions de transcription et bonnes pratiques

Qui est qui ?

Ecosystème administratif et humain, écosystème de projets

Département du Minutier central des notaires de Paris
En appui : département de la maîtrise d'ouvrage du système d'information - (auj. département du système d'information)

Agents scientifiques (archivistes, archivistes paléographes, ingénieurs de recherche, informaticien), agents techniques, stagiaires.

Département de l'innovation numérique (auj. département du numérique pour la transformation des politiques culturelles et de l'administration des données) du service du numérique (SNUM)

Initie, coordonne et accompagne la politique culturelle numérique de l'Etat

convention-cadre Culture- Inria (2016)

Equipe de recherche ALMAnaCH (*Automatic Language Modelling and Analysis & Computational Humanities*)

Chercheurs, ingénieurs recherche et développement, doctorants, post-doctorants et stagiaires.

Conventions-cadres : phasage et participations

Inria : recrutement, encadrement et formation de stagiaires, développeurs, ingénieurs R&D, infrastructure
MIC et AN : participation financière versée à Inria (AN : 20 %)

Phase 1	4 mois	15000 €
Phase 2	12 mois	65000 €
Phase 3 	12 mois	

(09/2020 : capacité de calcul Traces6, augmentation des discontinuités côté AN liées aux adaptations nombreuses des mesures sanitaires de prévention)

Phase 3 bis 	12 mois (avenant Covid)	65000 €

ARCHIVES NATIONALES

- **Aurélia Rostaing** (3 phases) - archiviste-paléographe, responsable du pôle instruments de recherche, cheffe de projet métier
- **Marie-Françoise Limon-Bonnet** (3 phases) - archiviste-paléographe, responsable du DMC
- **Danis Habib** (3 phases) - chargé d'études documentaires, formateur EAD
- **Marc Durand** - chargé d'études documentaires (phases 2 et 3)
- **Benjamin Davy** - agent technique (phases 2 et 3)
- **Nathalie Denis** - agent technique (phase 3)
- **Anna Cheru** - stagiaire (phase 2)
- **Florentine Ménager** - stagiaire (phase 3)
- **Marie-Véronique Vaillant** - attachée d'administration centrale (phase 3)
- **Pierre Bureau** - secrétaire de documentation (phase 3)
- **Virginie Grégoire** - secrétaire de documentation (phase 2)
- **Stéphane Grégoire** - agent technique (phases 2 et 3)
- **Nicolas Gros** - agent technique (phase 3)
- **Franck Beltrami** - agent technique (phase 3)
- **Charlotte Dridi** - agent technique (phase 3)
- **Fatima Reghida** - agent technique (phase 3)
- **Gaetano Piraino** (3 phases, surtout 1 et 2) – ingénieur de recherche
- **Frédéric Zamarreno** (3 phases, surtout 1 et 2) – responsable du DMOASI (auj. DSI)

Inria

- **Alix Chagué** (phases 2 et 3) - ingénieure R&D, cheffe de projet
- **Laurent Romary** - directeur de recherche, aujourd'hui directeur à la culture et à l'information scientifique
- **Yves Constantin Tadjo Takianpi** (phase 3) - ingénieur R&D
- **Éric Villemonte de la Clergerie** - chargé de recherche
- **Marie-Laurence Bonhomme** (phase 1) – stagiaire de master 2 TNAH - École nationale des chartes
- **Charles Riondet** (phases 1 et 2) – ingénieur R&D
- **Marie Puren** (phase 1) – ingénieure R&D
- **Lionel Tadjou** (phases 2 et 3) – ingénieur R&D
- **Lucas Terriel** (phase 3) – stagiaire de master 2 TNAH - École nationale des chartes
- **Hugo Scheithauer** (phase 3) – stagiaire de master 2 TNAH - École nationale des chartes

Equipe Scripta PSL - Histoire et pratiques de l’écrit

- **Daniel Stökl Ben Ezra** - directeur d’études (EPHE)
- **Peter Stokes** - directeur d’études (EPHE)
- **Marc Bui** - directeur d’études cumulant (EPHE)
- **Benjamin Kiessling** - développeur
- **Robin Tissot** - développeur
- **El Hassane Gargem** - développeur (phase 3)

En partenariat, l’équipe Teklia :

- **Bastien Abadie** - directeur technique (phase 3)
- **Eva Bardou** (phase 3)

Environnement technique et scientifique à Inria

Développement, manipulation de données…

GitLab

Accessibilité et science ouverte avec, par exemple, Gitlab

Formation à la recherche

Inria

eScriptorium (Scripta, PSL)

Infrastructure réseau

Porter des valeurs communes…

ARCHIVES NATIONALES

- Service aux publics
- Normalisation
- Optimisation des tâches (intégration de l’automatisation et de l’IA dans les processus métiers)

- Interopérabilité
- Science ouverte
- Standards de données

…et partager des objectifs communs

- Philosophie de synergie et de partage
- Mutualisation des outils et des infrastructures
- Passage à l’échelle
- Logiciels libres pour assurer la pérennité et l'interopérabilité des données
- Ouverture et réutilisabilité des outils et des données
- Recherche et veille
- Partage des compétences
- Vulgarisation scientifique réciproque

Une association de compétences métier

ARCHIVES NATIONALES

- Paléographie analogique (et numérique)
- Diplomatique (notariale) et histoire des fonds (originaux et leurs dématérialisations)
- Maîtrise des standards de description de données archivistiques (ISAD-G, XML-EAD)
- Recherche historique
- Contacts avec la recherche exploitant les fonds notariaux, ou en histoire notariale.

- Gestion de projet R&D
- Développement informatique
- Maîtrise et connaissance des outils état de l'art, principalement dans le domaine du traitement automatique des langues, mais aussi dans les domaines qui y sont liés
- Maîtrise des standards de données (ALTO, PAGE XML, XML-TEI, etc.)
- Science ouverte
- Recherche

Quel rôle pour les ingénieurs d'ALMAnaCH au sein de LectAuRep ?

- Prendre en main et proposer les outils appropriés aux tâches envisagées, sans réinventer la roue, et quand cela est nécessaire, développer des solutions spécifiques.
- Assurer le transfert de compétences pour que la prise en main des outils soit assurée pour le DMC.
- Maîtrise de l'apprentissage machine, analyse et interprétation des métriques obtenues à partir de l'évaluation des modèles d'intelligence artificielle.
- Science des données : organiser la création des corpus d'entraînement, tracer les entraînements de modèles, maîtrise des standards de données.
- Veille scientifique et technique.

Les stages : une opportunité pour explorer

Marie-Laurence Bonhomme (2018)

→ Rapport exploratoire sur la segmentation automatique et la reconnaissance d'écriture manuscrite dans les répertoires des notaires.

Lucas Terriel (2020)

→ Création d'un format pivot TEI pour centraliser les métadonnées associées aux documents et celles générées durant l'étape de transcription automatique, et création d'un outil pour évaluer la qualité des transcriptions (KaMI)

Hugo Scheithauer (2021)

→ Exploration de la reconnaissance d'entités nommées dans les répertoires et des solutions de publication des répertoires transcrits (éditorialisation des données)

Transkribus®

TEI

tei Publisher
The Instant Publishing Toolbox

Utiliser un outil libre : le choix d’eScriptorium

- eScriptorium est un logiciel libre développé par l’ équipe Scripta (PSL).
- Au terme d’un état de l’art initial, ALMAnaCH préconisa son utilisation, choix partagé par le DMC.
- Il a donc été décidé d’utiliser un outil générique, non spécifique au projet.
- Cependant, ALMAnaCH a entretenu, et entretient encore, une collaboration étroite avec l’équipe Scripta dans le développement de l’application.

eScriptorium : un outil de transcription et de gestion de projet

- Permet le travail asynchrone
- Planification de campagnes de transcription
- Une interface web rapide à prendre en main qui facilite la collaboration entre les agents.

→ Mais l’utilisation d’un service disponible sur le web sous-entend l’accès à une infrastructure informatique robuste et adaptée.

Mutualiser les outils : infrastructures

→ Mettre à disposition de la communauté
un outil facile à prendre en main
pour transcrire des documents écrits

- 2019 : machine virtuelle, sans carte graphique.
- 2020 (septembre) : serveur Traces6, 2 cartes graphiques, plus de stockage.
- 2022 : CREMMA, 2 cartes graphiques supplémentaires, plus de stockage, plus de mémoire vive.

Mutualiser les outils : infrastructures

→ Sharedocs :
sauvegarde des données,
partage de documents
entre les différents acteurs du projet,
→ expérimentations avec IIIF.

International
Image
Interoperability
Framework

Communiquer sur le projet

L'ingénieur a pour mission de vulgariser son travail pour que tous les acteurs, aussi bien internes au projet qu'externes, comprennent les enjeux et les processus. L'archiviste a en ligne de mire l'intégration future de certains outils dans son SIA (si possible), voire dans les SIA proposés par les prestataires de solutions métier, et une convergence quand elle est possible (GLAM, ESR, prestataires).

Carnet de recherche lancé et maintenu par Alix Chagué, et contributions internes à l'équipe ALMAnaCH

hypotheses

https://lectaurep.hypotheses.org/

Rôle important des interventions publiques

AI4LAM

BnF

Liberté • Égalité • Fraternité
RÉPUBLIQUE FRANÇAISE
MINISTÈRE DE L'EUROPE ET DES AFFAIRES ÉTRANGÈRES

Liberté • Égalité • Fraternité
RÉPUBLIQUE FRANÇAISE
Ministère Culture

#dhnord humanités numériques

MATE-SHS

GitHub

GitLab

HTR United

Lab Intelligence Artificielle

etalab gouv.fr

LectAuRep : résultats en prime

Création ou exploitation d'outils génériques
- Outil d'évaluation du désaccord interannotateurs (Alix Chagué)
- KaMI (Lucas Terriel) > un outil d'évaluation agnostique, exploitable quel que soit le logiciel d'HTR ayant produit les données, dans un contexte de banc d'essai de logiciel, de publication de données, de marché public ; calibrable selon la nature des corpus (riches ou pas en chiffres, en signes diacritiques…)
- Un prototype TEI Publisher pour le projet LectAuRep (Hugo Scheithauer)

Equipe Scripta-PSL
- Remontées de retours utilisateurs et de cas d'usage
- Développement de fonctionnalités pour eScriptorium (étiquetage de documents, tableau de bord…)
- Création de documentation pour eScriptorium (tutoriel d'Alix Chagué)

GLAM & chercheurs
- Pipeline d'export de données de Transkribus, test d'import dans Kraken (Aspyre, Alix Chagué)
- Partage d'expérience avec des porteurs de projet et des groupes de travail (labIA de la BnF, datadrink d'Etalab, AI4LAM, CREMMALab, réseau AEOLIAN…)
- Communications et publications scientifiques (Hypothèses)
- Partage d'un corpus de données très spécifiques en langue française (langage technique, elliptique et fortement abrégé, riche en entités nommées)

Conclusion

- **La réalité de l'organisation du travail**
  - Une collaboration hebdomadaire (phase 3 bis) dans une confiance mutuelle entre des cheffes de projet "voies d'aiguillage" (arbitrages de premier niveau avant celui, final, de Scripta-PSL)
  - Partie du service côté Inria : serveur Traces6, entraînements en lignes de commande, analyse des métriques, mise au point de modèles.

- **Aspects pratiques et échanges nécessaires à l'avancée de la recherche**
  - Courriels, puis recours aux Gitlab ALMAnaCH et Scripta (suivi des entraînements, remontée de bugs, propositions de fonctionnalités à Scripta - phase 3 bis)

- **Ecosystème du mode projet**
  - Pas de comité de pilotage mais des réunions plénières de lancement (phases 1 et 2), une réunion de prérapport (phase 3 bis) ; 2 rapports.
  - Une feuille de route très large et une grande liberté de manœuvre
  - Une organisation très souple, rythmée (phase 3 bis) par les réunions ALMAnaCH-Scripta

- **La collaboration : utopie ou réalité ?**
  - **La collaboration sur ce projet n'a pas relevé de l'utopie. Au regard des besoins ce fut une nécessité, tant du point de vue des compétences que des infrastructures à mobiliser dans le cadre de cette exploration.**

Quelques retours de la collaboration sur le terrain

Méthodologie, priorités et stratégie : des solutions assumées ou négociées
> Abandon de l'idée de fork et participation aux réunions de développement de Scripta (Inria, phase 3 bis)
> Rappel de l'objectif de transcrire le random set (Inria)
>> Décision de sécuriser le projet en traitant les CM en l'absence d'un modèle de segmentation performant pour le XXe s. (AN)

Un ou plusieurs annotateurs par page ?
Mise en relief du désaccord interannotateurs (Inria-Alix Chagué) vs. gestion optimisée des opérateurs disponibles (AN) >> KaMI (Inria-Lucas Terriel)

Transcription parfaite ou imparfaite ?
Prise de conscience épistémologique d'un "biais paléographique et diplomatique"
> du stemma au modèle générique
> difficulté de traitement nettement supérieure de certaines écritures du XXe s. par rapport à d'autres d'Ancien Régime
> purisme paléographique vs. efficacité de transcriptions fautives pour un premier modèle à affiner
> importance de l'effet de corpus (au sens diplomatique) pour l'HTR (AN)

De l'intérêt de la TEI pour éditorialiser et fouiller des données d'HTR
> TEI Publisher (Inria-Hugo Scheithauer)

S’il fallait recommencer ?

> Intégrer Scripta-PSL à la convention-cadre
> Rédiger en amont des conventions écrites de segmentation et de transcription et des supports d’accompagnement
> Mieux quantifier et qualifier le corpus cible, mieux calibrer l’échantillonnage des corpus de tests (avec Sparnatural ?) et des corpus d’entraînement

Des regrets ?

> Ne pas avoir rédigé un plan de gestion de données d’HTR (hors objectif)
> Ne pas (encore) avoir rédigé de diagnostic pour un passage à l’échelle (ingénierie de projet, infrastructures, logistique collaborative)

Des satisfactions ?

> Avoir pris en charge un patrimoine numérique mixte (NB d’après microfilm et couleur d’après originaux)
> Mettre à disposition un embryon de corpus francophone (archivistico-historique plus que littéraire ou en langage naturel) pour du TAL (précorrection automatique), de la REN sur des entités nommées d’HTR bruité
> Avoir impulsé, et maintenu en période de crise, le premier projet d’HTR issu d’un service d’archives en France
> Avoir partagé des retours d’expérience auprès des réseaux francophone et anglophone du patrimoine écrit

Merci pour votre attention !

Des questions ?

aurelia.rostaing@culture.gouv.fr
hugo.scheithauer@inria.fr

Carnet de recherche : https://lectaurep.hypotheses.org/

Gitlab : https://gitlab.inria.fr/almanach/lectaurep
Github : https://github.com/lectaurep

Données déposées sur HTR-United :

- https://github.com/HTR-United/lectaurep-mariages-et-divorces
- https://github.com/HTR-United/lectaurep-bronod
- https://github.com/HTR-United/lectaurep-repertoires

ARCHIVES NATIONALES

Inria