Archives de hier et de demain

Laurent Romary
Inria & DIM MAP Dopamines

Présentation personnelle rapide

- Recherche en informatique : ancrée sur une démarche pluri-disciplinaire
  - Traitement automatique des langues
  - Modélisation de données en sciences humaines (langue, documents)
- Du modèle au standard : participation à la normalisation internationale
  - *Text Encoding Initiative* : consortium de référence pour la représentation numérique de textes
  - Comité TC 37 de l'ISO (*Langue et terminologie*)
- Contribution aux infrastructures européennes
  - Initiateur (depuis 2006) et directeur (2014-2018) de l'infrastructure DARIAH ERIC
- Une implication (obstinée) dans la science ouverte
  - CNRS, Société Max Planck, Inria; membre du SPSO (Secrétariat Permanent pour la Sciences Ouverte)
  - Déploiement de HAL, développement des centres de ressources numériques (bases des consortiums Huma-Num), obligation de dépôt dans HAL à Inria

Archives et recherches en numérique

- Gérer le tournant numérique
  - Accélération du processus de numérisation de documents anciens, couplé à une production de données naturellement numériques (*born-digital*)
  - Présentation de quelques projets couplant recherche et infrastructure à différentes étapes de ces développements
- Présentations de différents projets liés au lien entre archives et numérique
  - Partie 1: travail sur les matériaux anciens (ou moins anciens)
    - Intégration de données hétérogènes en EAD dans le projet EHRI
    - Extraction d'information à partir d'actes notariés (Archives Nationales)
    - Numérisation et structure: donner du sens
      - Grobid et entity-fishing
  - Partie 2: archives des données de la recherche
    - Travailler à la traçabilité des données
    - Scénario E-RIHS Enquête Iperion CH
    - Pistes à suivre: Standard (cf. Vienne) – TEI – SSK; Charte – DMP; Sciences Call – questionnaire

Un peu de contexte: TEI et EAD

- TEI – Text Encoding Intitiative
  - Directives pour le codage de données textuelles en XML
  - Initié en 1987 – édition courante P5
  - Standard ouvert (GitHub), maintenance réactive (2 releases par an)
  - ODD (One Document Does it all): language de spécification de la TEI
- EAD - Encoded Archival Description
  - Encodage des instruments de recherche archivistiques en XML (importance du niveau collection vs. objet)
  - Initié en 1993, EAD 1.0 (1998) … EAD3 (2015)
  - Maintenance par la *Society of American Archivists* et la Bibliothèque du Congrès
  - Inspiré de la TEI (cf. D. Pitti)
  - Permet un alignement sur la norme ISAD(G)

Remerciements à Veerle Vanden Daelen, Charles Riondet et toute l'équipe du projet EHRI

INTÉGRATION DE DONNÉES HÉTÉROGÈNES ISSUES DES ARCHIVES DE L'HOLOCAUSTE

EHRI: Basic information

- EHRI's second phase (2015-2019) as an EU financed project with a total budget of almost 8 mio €
- 24 partner institutions from 17 countries: Research institutions, archives and e-science specialists
- EHRI's goal: Support research into the Holocaust and help networking of Holocaust researchers and archives

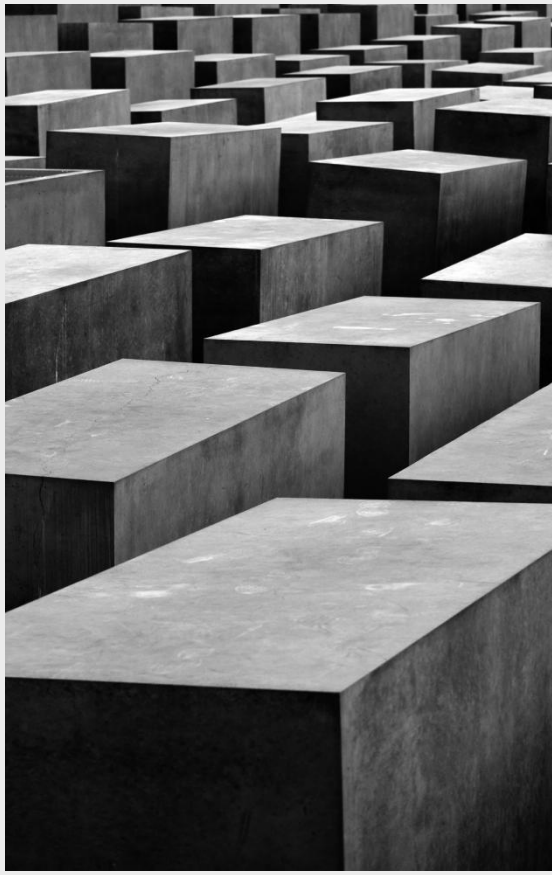
EHRI: Partner institutions

- NIOD, Institute for War, Holocaust and Genocide Studies (Amsterdam): Overall project coordination
- Yad Vashem (Jerusalem)
- CEGESOMA (Brussels)
- King's College (London)
- Institute for Contemporary History (Munich)
- Jewish Museum in Prague
- DANS (Den Haag)
- Wiener Library (London)
- Vienna Wiesenthal Institute for Holocaust Studies
- ŻIH (Warsaw)
- Mémorial de la Shoah (Paris)
- International Tracing Service (Arolsen)
- USHMM (Washington D.C.)
- Bundesarchiv (Berlin / Koblenz)
- Elie Wiesel National Institute for the Study of the Holocaust in Romania (Bucharest)
- Hungarian Jewish Archives (Budapest)
- Vilna Gaon State Jewish Museum
- Dokumentačné stredisko holokaustu (Bratislava)
- Foundation Jewish Contemporary Documentation Center CDEC (Milan)
- The Jewish Museum of Greece (GR)
- Ontotext (Sofia)
- INRIA (Le Chesnay)
- Stowarzyszenie Centrum Badań nad Zagładą Żydów (Warsaw)
- Kazerne Dossin: Memorial, Museum and Documentation Centre on Holocaust and Human Rights (Mechelen)

EHRI Aims

The main objective of EHRI is to support the Holocaust research community by

1. **integrating** information on key archival **collections** and **institutions** into an online portal
2. **encouraging** collaborative Holocaust **research** and investigating new methodologies



www.ehri-project.eu

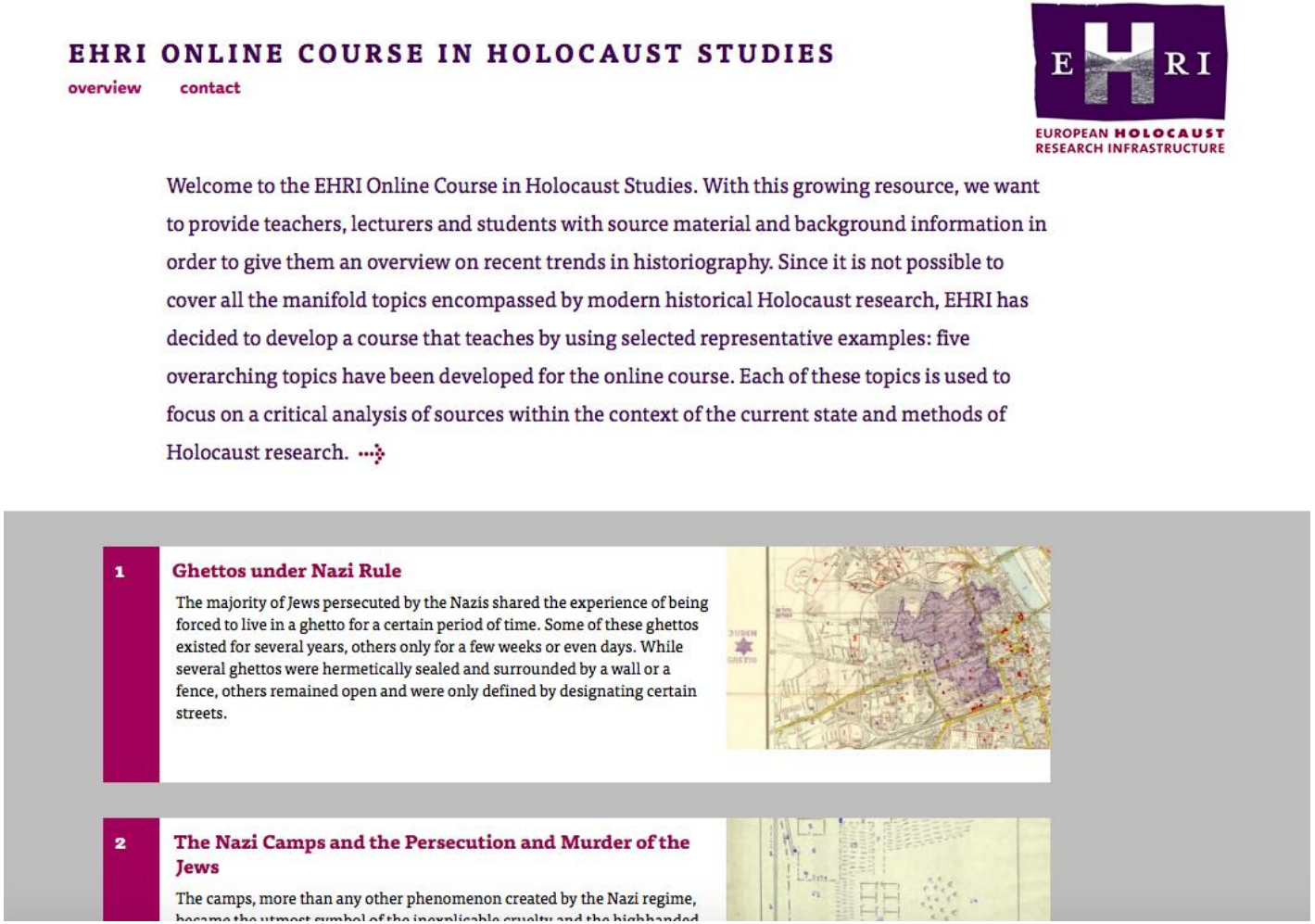
Why EHRI?

- **Fragmentation and dispersal of archival sources**
  - Geographical scope Holocaust
  - Attempts to destroy the evidence
  - Migration of Holocaust survivors
  - Multiple documentation projects after the war
- **Internationalization Holocaust research**
  - Holocaust in Eastern Europe
  - New levels of collaborative research
- **New opportunities for digital research**

Seminars, Online Courses, Document Blog

- **Methodological Seminars**, including an upcoming seminar for **conservationists** working on Holocaust material
- Next to the existing **Online Course** http://training.ehri-project.eu , e.g. **interactive courses** and Bundesarchiv-written course on **German Archivistics (Aktenkunde)**
- **Workshops** on specific topics, **Research Guides** (e.g. on **Theresienstadt,** https://portal.ehri-project.eu/guides/terezin) and upcoming **online editions** as resources
- The **EHRI Document Blog** (https://blog.ehri-project.eu/)

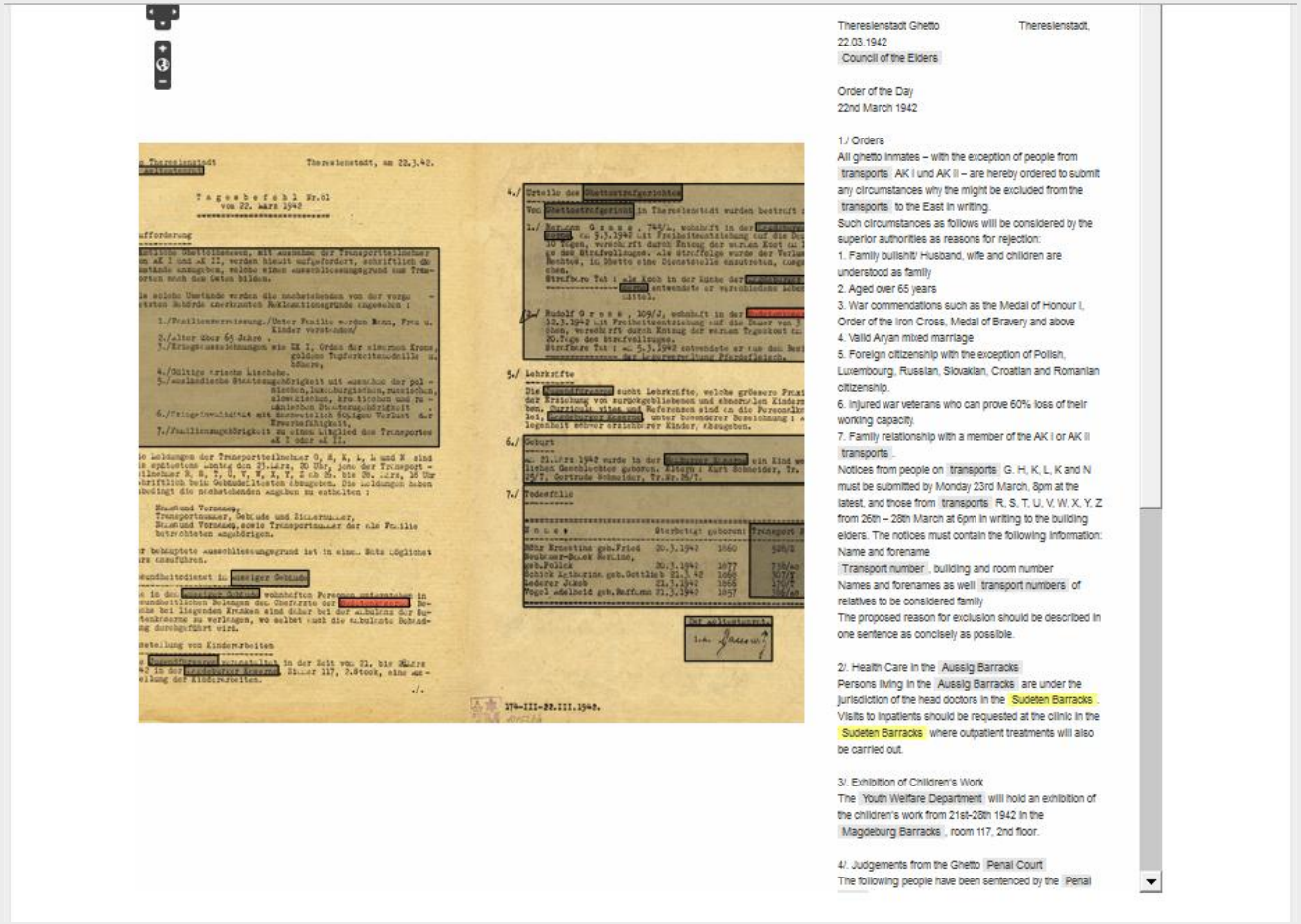
EHRI online courses - https://training.ehri-project.eu



unguided online course with 6 units; tutored online course with 17 chapters

EHRI Document Blog - https://blog.ehri-project.eu

Innovative platform for the interpretation, visualisation and contextualisation of Holocaust sources



Theresienstadt Ghetto Theresienstadt,
22.03.1942
Council of the Elders

Order of the Day
22nd March 1942

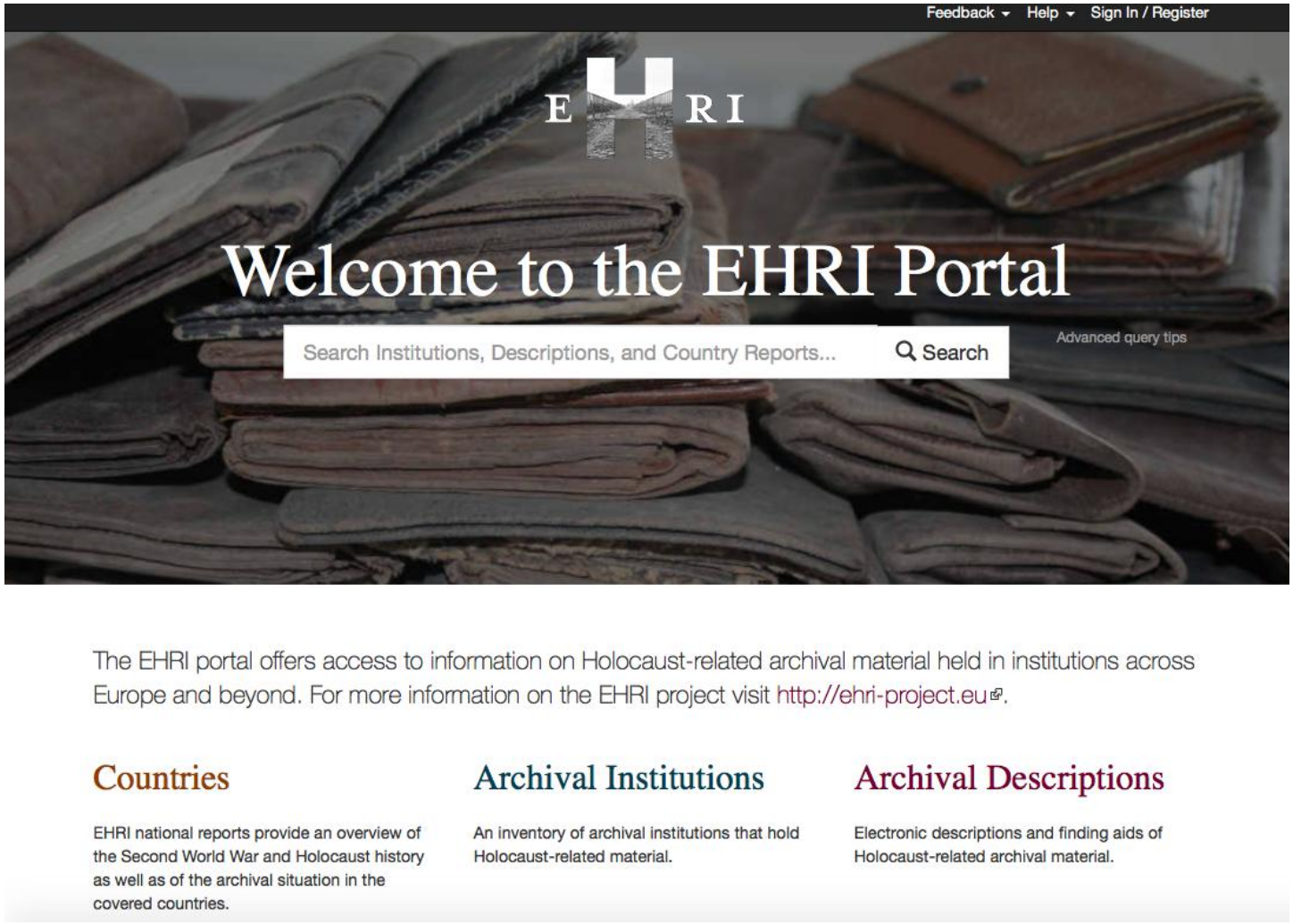
1./ Orders
All ghetto inmates – with the exception of people from transports AK I und AK II – are hereby ordered to submit any circumstances why the might be excluded from the transports to the East in writing.
Such circumstances as follows will be considered by the superior authorities as reasons for rejection:
1. Family bullshit/ Husband, wife and children are understood as family
2. Aged over 65 years
3. War commendations such as the Medal of Honour I, Order of the Iron Cross, Medal of Bravery and above
4. Valid Arjan mixed marriage
5. Foreign citizenship with the exception of Polish, Luxembourg, Russian, Slovakian, Croatian and Romanian citizenship.
6. Injured war veterans who can prove 60% loss of their working capacity
7. Family relationship with a member of the AK I or AK II transports.
Notices from people on transports G, H, K, L, K and N must be submitted by Monday 23rd March, 8pm at the latest, and those from transports R, S, T, U, V, W, X, Y, Z from 26th – 28th March at 6pm in writing to the building elders. The notices must contain the following information:
Name and forename
Transport number, building and room number
Names and forenames as well transport numbers of relatives to be considered family
The proposed reason for exclusion should be described in one sentence as concisely as possible.

2/. Health Care in the Aussig Barracks
Persons living in the Aussig Barracks are under the jurisdiction of the head doctors in the Sudeten Barracks. Visits to inpatients should be requested at the clinic in the Sudeten Barracks where outpatient treatments will also be carried out.

3/. Exhibition of Children's Work
The Youth Welfare Department will hold an exhibition of the children's work from 21st-28th 1942 in the Magdeburg Barracks, room 117, 2nd floor.

4/. Judgements from the Ghetto Penal Court
The following people have been sentenced by the Penal

and: EHRI Portal - https://portal.ehri-project.eu/



57 country reports, >1,900 descriptions of institutions, >230,000 archival descriptions

The EHRI Portal

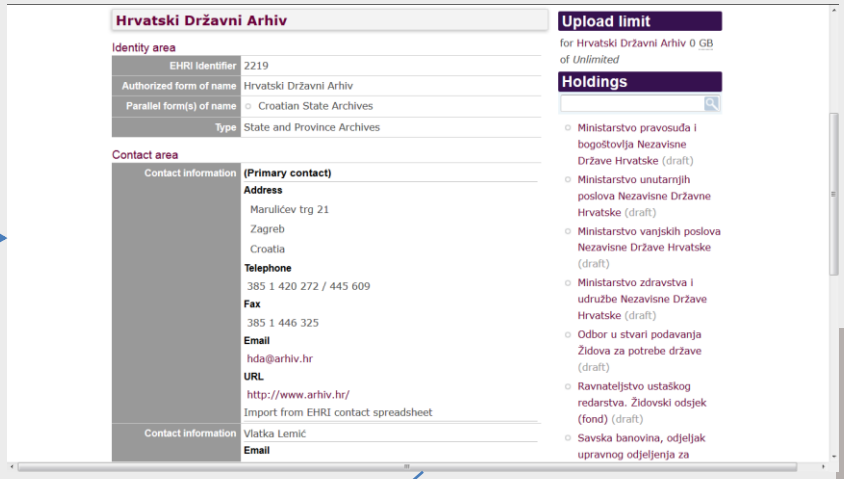
The goal is to expand the **online inventory of institutions and collections** pertaining to the Holocaust in Europe, Israel and beyond

- Making sources visible in a systematic fashion in order to counteract the fragmentation of the sources
- Reveals interconnections (e.g. through a multilingual thesaurus with approx. 5470 terms; collation of authority files; relationships between originals and copies)
- EHRI focuses on **collection descriptions** – it is not aiming to be a "scan depository", nor does it aim to be a complete public database on the (often privacy sensitive) file or document level (although we will take those descriptions if we can have them)
- The ultimate goal is to connect archives and users (mutually useful -> e.g. expert user feedback)

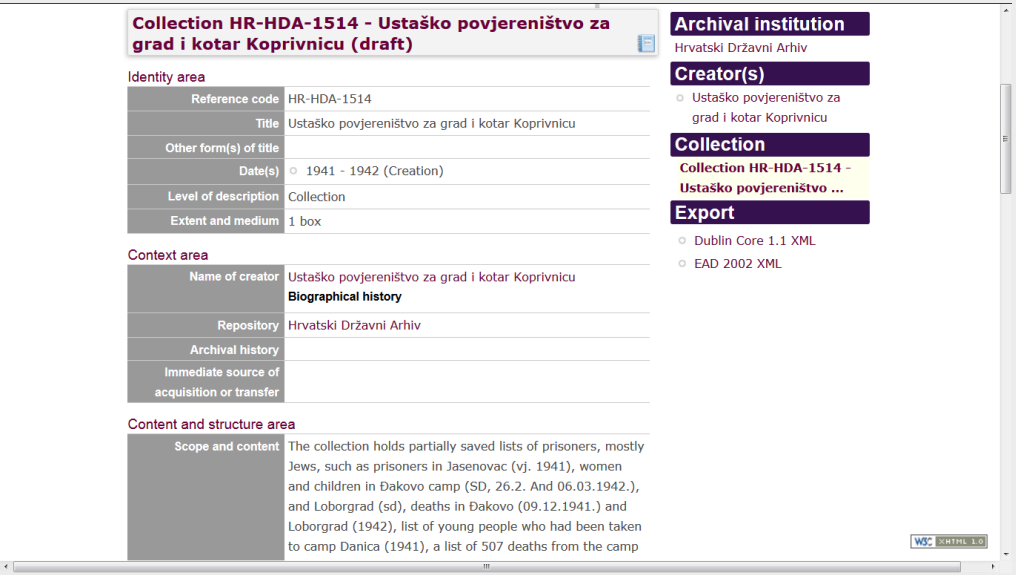
EHRI Database

Country reports (57 countries)

Entry on the individual archive (over 1,900)

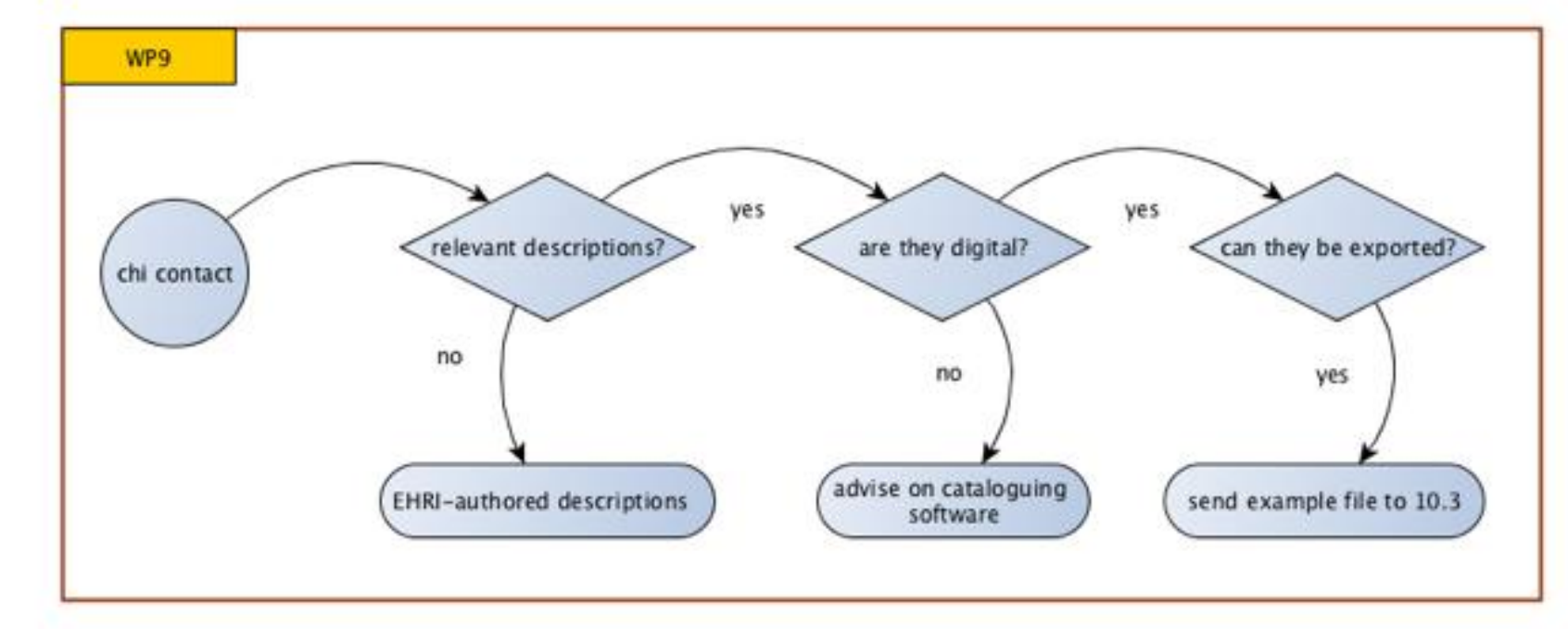


Individual entries (collections / units) (tens of thousands)

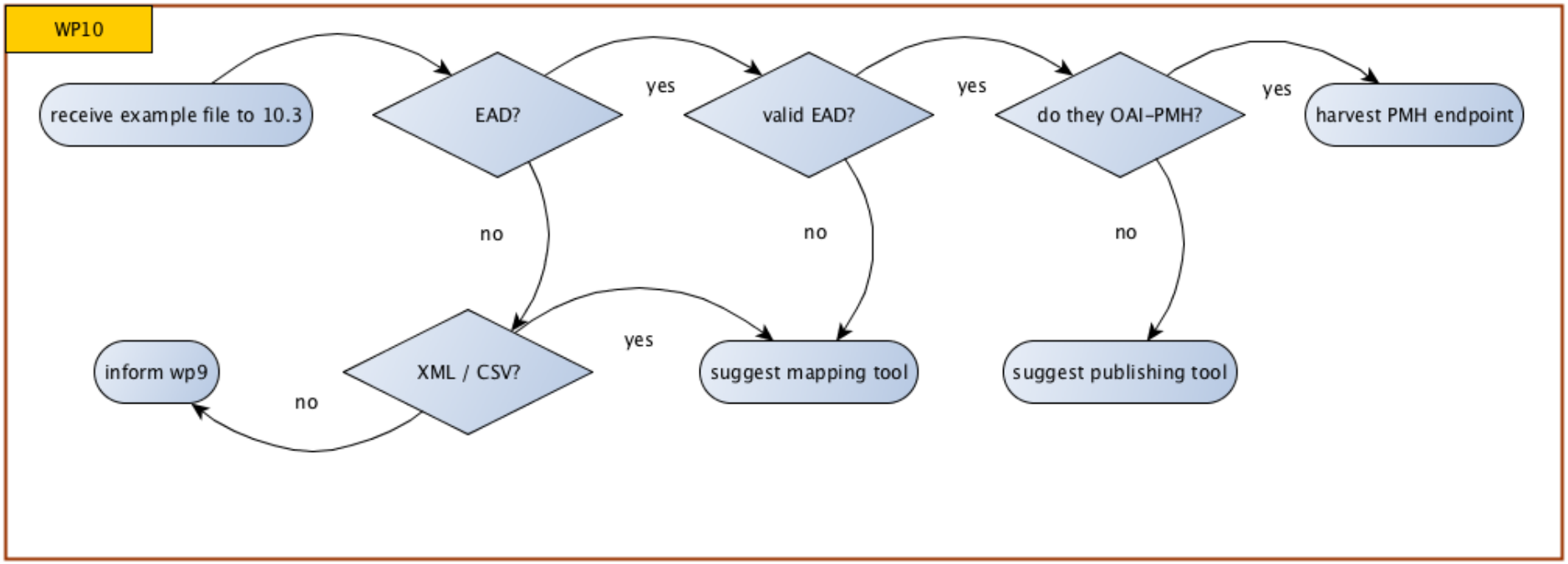


[EHRI Portal]

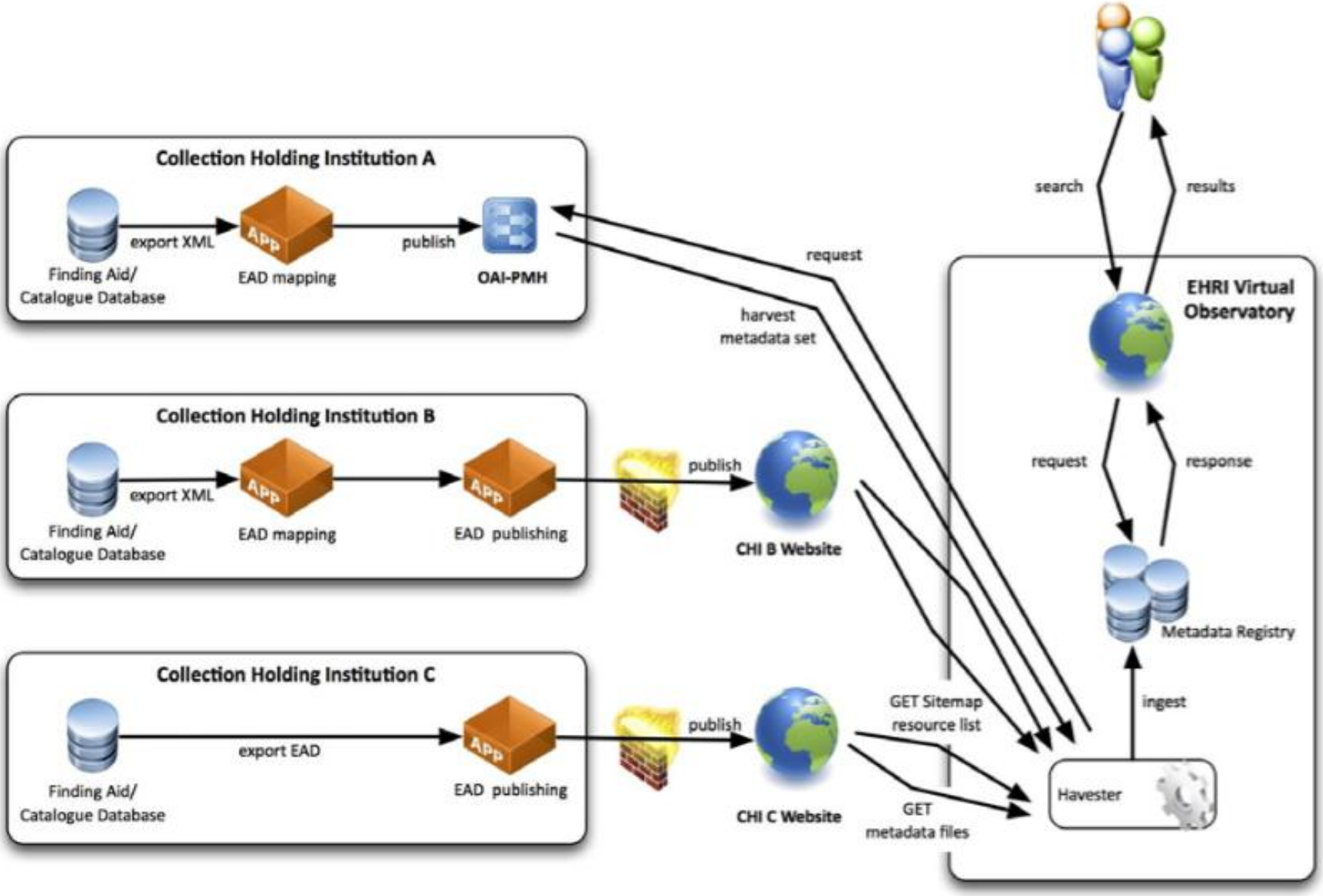
Integration of collection descriptions (I)



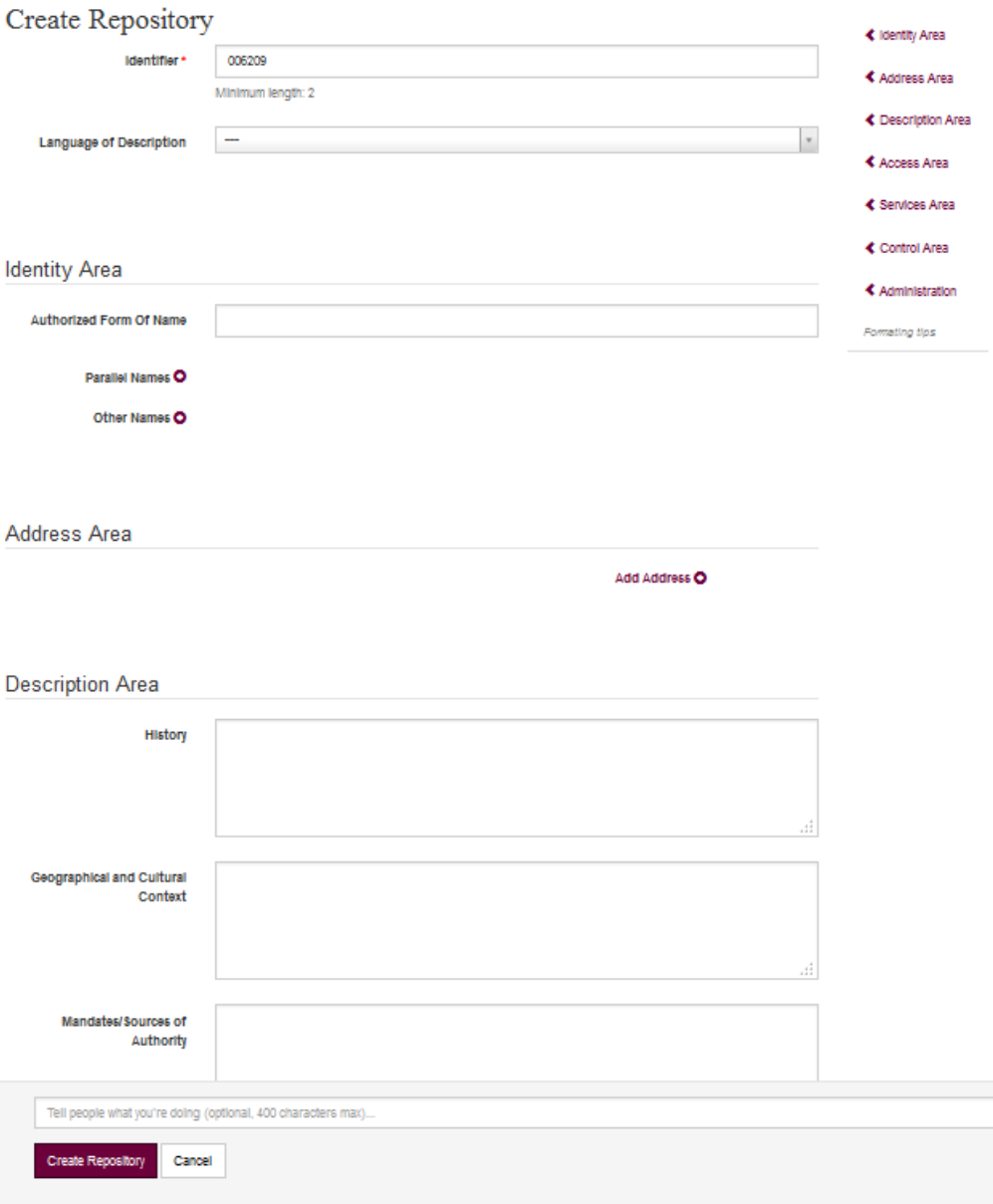
Integration of collection descriptions (II)



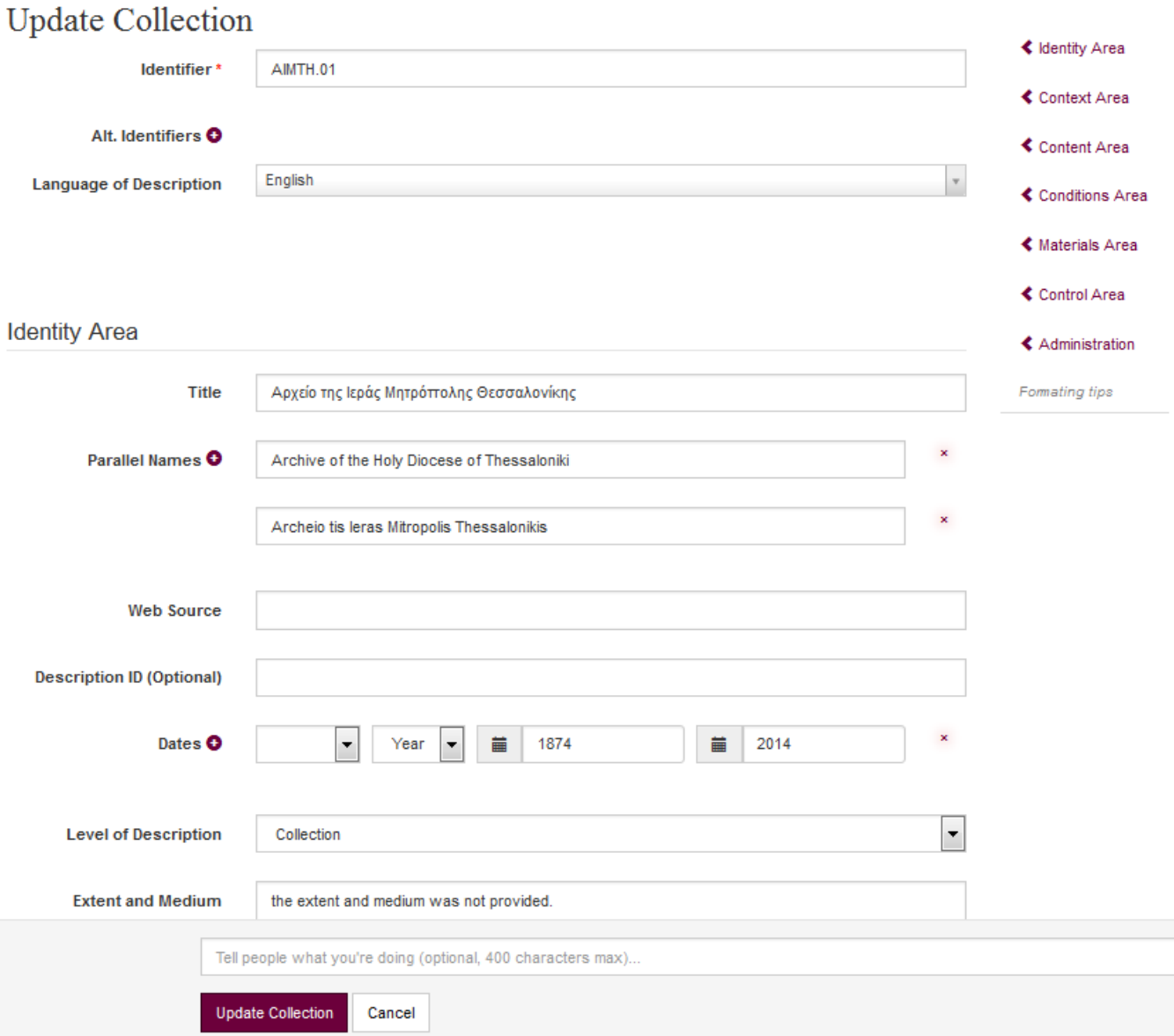
Data Integration via mapping & publishing tool



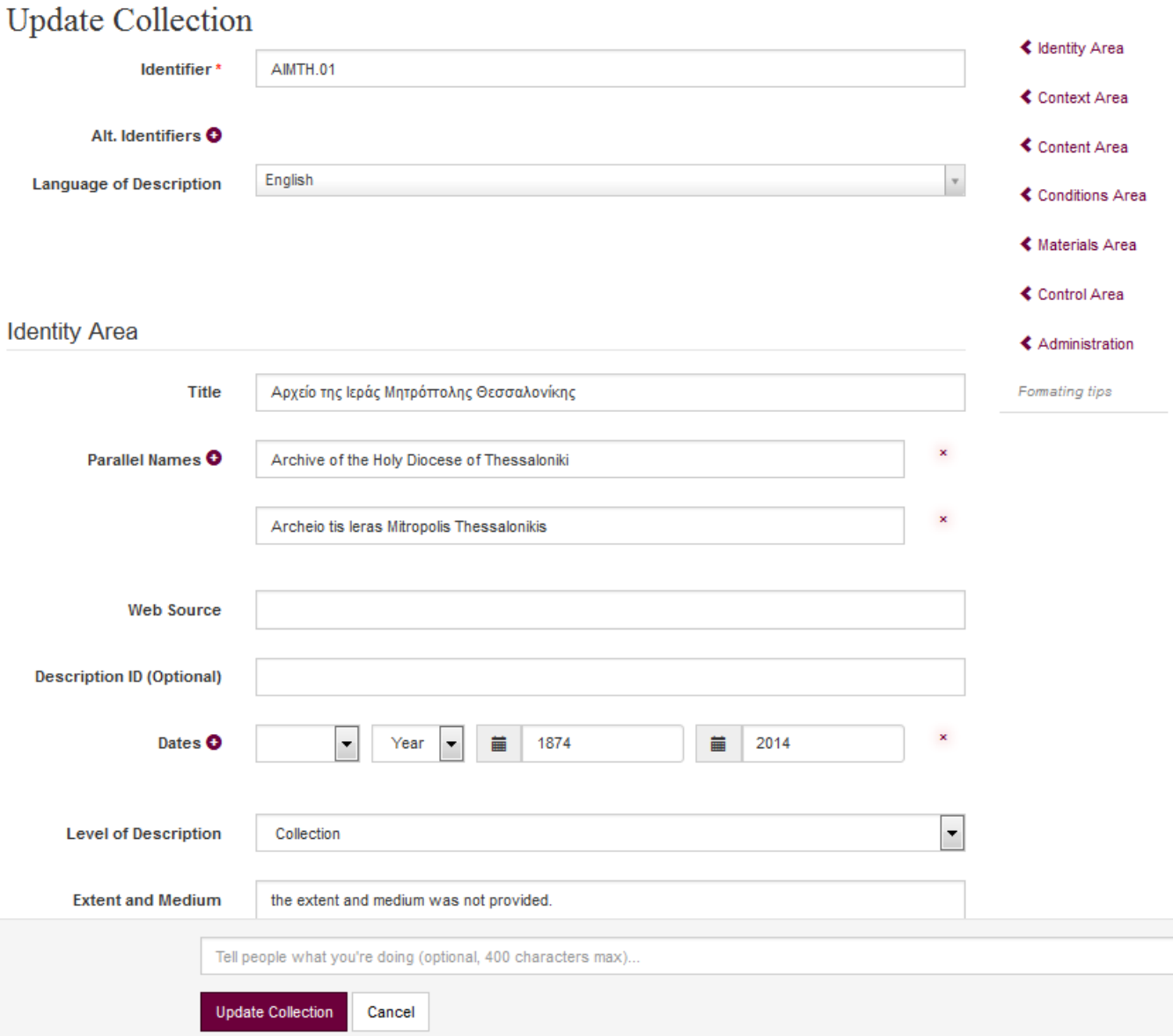
Manual repository descriptions



Creating collection or child item descriptions



Manual collection updates



The EHRI EAD customisation model

Focus on the role and the value of community standards, like EAD and TEI in data integration, enrichment, sharing lifecycle.

→ Interoperability

→ Enrichment

→ Favour new uses

EAD, what’s wrong with it?

Since the beginning of EAD, its “permissiveness” is seen as a challenge or a weakness :

Shaw 2001:

"A more prescriptive descriptive standard (...) will greatly enhance the potential for machine processing of finding aids across repositories”.

Bunn 2013:

- Need to "Draw a distinction between information exchange and archival description”
- "Still no standard for archival description”

EAD, what's wrong with it?

EAD doesn't have all the means to handle its flexibility.

How to preserve meaning and content?

→ Exchange and connect archival data together, and together with others resources available online

RiC (Records in Context) → New paradigm? Definitive solution?

EHRI and standards

→ EAD2002 is the pivot format for automatic ingestion of archival descriptions.

→ Ingestion of data in many formats

- → EAD1, Dublin Core, home made formats
- → EAD 2002 with very different encoding guidelines

→ EHRI has its own specific description rules

Project-specific EAD customizations

What is customization?

- Narrowing EAD countless encoding possibilities
- Adding quality checks
- Content-oriented rules
- **Not modifying the schema**

TEI long lasting experience // One document does it all (ODD)

Maintain both the technical and editorial content within a single framework.

→ Schema fragments, prose documentation and reference documentation in a single document

ODD allows for total flexibility to model specific subsets or extensions of the described format.

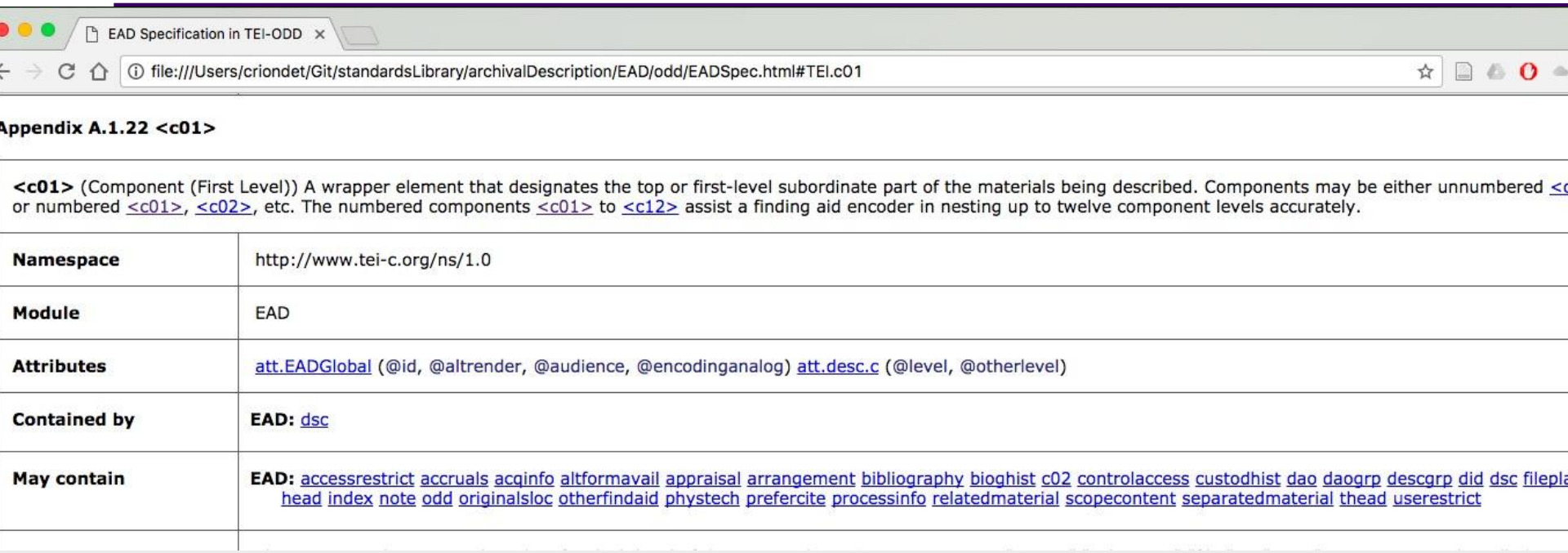
TEI long lasting experience // One document does it all (ODD)

- Straightforward way to customize an XML format according to specific practices and document this customization
- Possible to describe and generate any schema and its documentation with XSL transformations
  - TEI Roma service
  - Saxon or other engine

Defining an XML element with ODD

```
<elementSpec ident="c01" module="EAD">
    <gloss>Component (First Level)</gloss>
    <desc>A wrapper element that designates the top or first-level subordinate
        part of the materials being described. Components may be either
        unnumbered <gi>c</gi> or numbered <gi>c01</gi>, <gi>c02</gi>, etc. The
        numbered components <gi>c01</gi> to <gi>c12</gi> assist a finding aid
        encoder in nesting up to twelve component levels accurately.</desc>
    <classes>
        <memberOf key="att.EADGlobal"/>
        <memberOf key="att.desc.c"/>
    </classes>
    <content>
        <rng:optional>
            <rng:ref name="head"/>
        </rng:optional>
        <rng:ref name="did"/>
        <rng:zeroOrMore>
            <rng:ref name="model.desc.full"/>
```

Generate schema and documentation




```
<define name="c01">
   <element name="c01">
      <a:documentation xmlns:a="http://relaxng.org/ns/compatibility/annotations/1.0">(Component
         (First Level)) A wrapper element that designates the top or first-level subordinate part
         of the materials being described. Components may be either unnumbered c or numbered c01,
         c02, etc. The numbered components c01 to c12 assist a finding aid encoder in nesting up
         to twelve component levels accurately.</a:documentation>
      <optional>
         <ref name="head"/>
      </optional>
      <ref name="did"/>
      <zeroOrMore>
         <ref name="model.desc.full"/>
      </zeroOrMore>
```

One document does it all and EAD

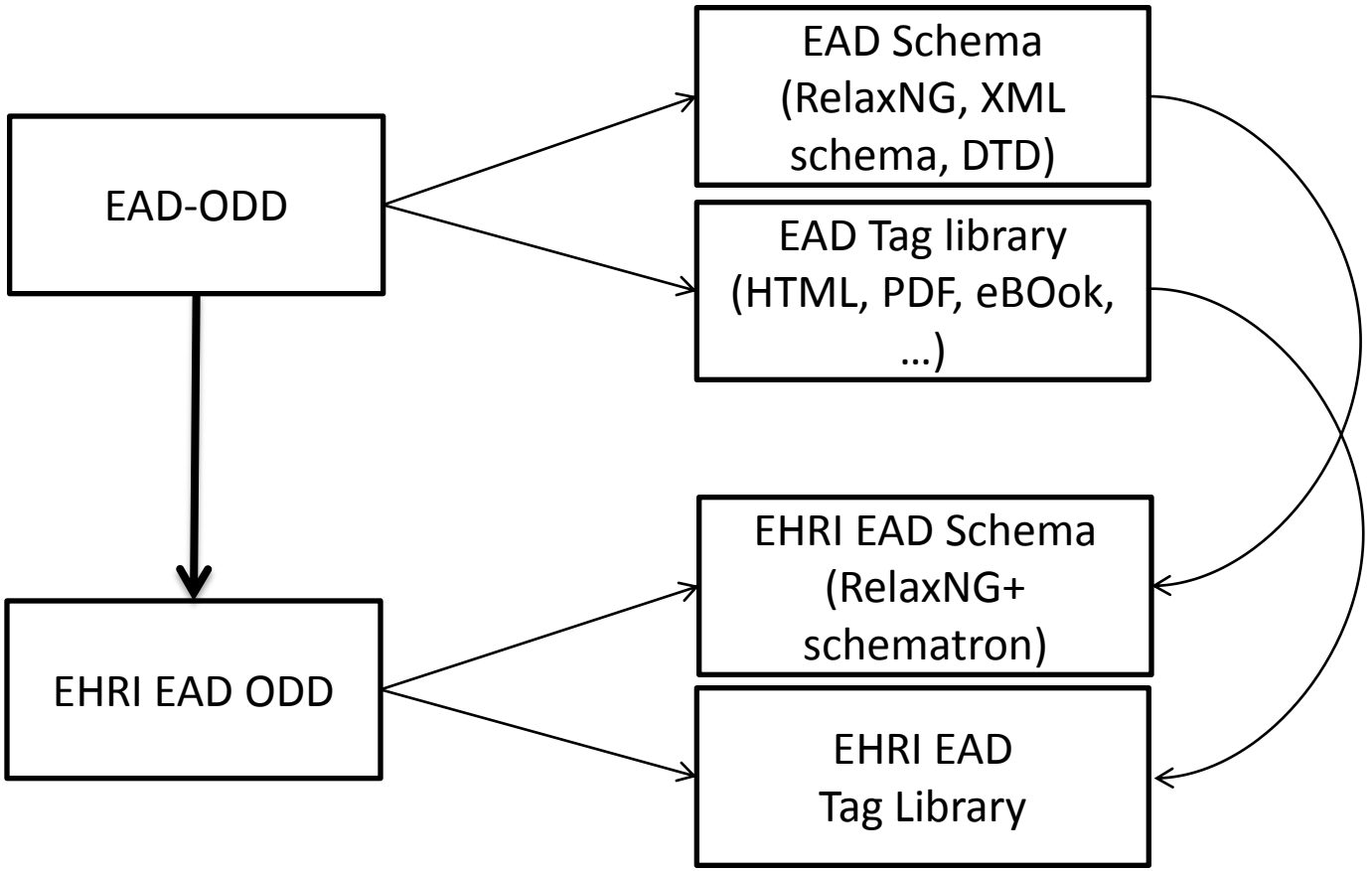
- We used ODD to cover EAD entirely
  - Official EAD schema (RelaxNG) : www.loc.gov/ead/ead.rng
  - Guidelines provided by the Library of Congress: http://loc.gov/ead

- Maintained by Parthenos (possibility to contribute and reuse) http://github.com/ParthenosWP4/standardsLibrary/blob/master/archivalDescription/EAD/ODD/EADSpec.xml

Flexible and customizable methodology

- Very precise content oriented rules based both on EHRI and CHI input data models
- Integrating the human readable documentation in the validation process → deepen the relationship between validation and documentation

Flexible and customizable methodology



Flexible and customizable methodology

- Use ODD to create specific EAD profiles
  - Each new EAD profile = new ODD
  - Inheritance to the master source + possibility to modify the elements that have a different behaviour
- Adding more specific rules to the core EAD schema
  - EAD schema (expressed in RelaxNG)
  - ISO Schematron rules

Schematron rules

- Emphasize EAD validation errors
- Align the descriptions with EHRI constraints
- Highlight some description elements that could be improved
- Sorted in categories (roles)
  - MUST: mandatory for import process
  - SHOULD: mandatory for description process, i.e. In terms if archival description. Not technically mandatory, but may cause comprehenstion issues
  - COULD: non mandatory rules. Enhance the general quality of the description, without any obligation. Pointing that informational element.

Schematron rules

```
<constraintSpec ident="labelDesirable" scheme="isoschematron" type="EHRI" mode="add">
  <desc><gi>unitdates</gi> COULD have a <att>label</att> attribute or an
      <att>encodinganalog</att> attribute, describing the type of date</desc>
  <constraint>
    <rule xmlns="http://purl.oclc.org/dsdl/schematron" context="ead:unitdate"
      see="&path;#EAD.unitdate"><assert xmlns="http://purl.oclc.org/dsdl/schematron"
        role="COULD" test="normalize-space(@label) or normalize-space(@encodinganalog)"
        >unitdates COULD have a label attribute or an encodinganalog attribute,
        describing the type of date</assert></rule>
  </constraint>
</constraintSpec>
```

Schematron rules

```
<constraintSpec ident="authfilenumberPossible" scheme="isoschematron" type="EHRI">
  <desc>Access points COULD be chosen in authority lists. The list is declared with a
      <att>source</att> attribute. The related id of this authority should be declared
    in an <att>authfilenumber</att> attribute. Note that EHRI provides URLs for
    vocabularies and authorities. Check the <ref target="http://ehri-project.eu">EHRI
      website</ref> for more information</desc>
  <constraint>
    <rule xmlns="http://purl.oclc.org/dsdl/schematron" context="ead:controlaccess"
      see="&path;#EAD.controlaccess">
      <assert xmlns="http://purl.oclc.org/dsdl/schematron" role="COULD"
        test=".[@authfilenumber and @source]">Access points COULD be chosen in an
        authority list. This list should be declared in a @source attribute. The related
        id of this authority should be declared in an @authfilenumber attribute.
      </assert>
    </rule>
  </constraint>
</constraintSpec>
```

Schematron rules

- Content normalisation (dates, codes, …)
- Required elements in EHRI (but not in EAD) → scopecontent (description of the content of the documents) for instance
- Improve descriptions quality. Not errors, but pieces of advice. In particular for content related elements (existence of copies of the material, bibliographic references, …)

Connect validation and mapping process to ad hoc documentation

A full description of the expected content (i.e. HTML “tag library”) is generated from the ODD file.

In the validation process, the documentation is served to the user in its context.

- EAD technical documentation
- EHRI technical documentation
- EHRI qualitative documentation

Connect validation and mapping process to ad hoc documentation

2.1.11. <archdesc>

<archdesc> (Archival Description) A wrapper element for the bulk of an EAD document instance, which describes the content, context, and extent of a body of archival materials, including administrative and supplemental information that facilitates use of the materials. Information is organized in unfolding, hierarchical levels that allow for a descriptive overview of the whole to be followed by more detailed views of the parts, designated by the element Description of Subordinate Components <dsc> . Data elements available at the <archdesc> level are repeated at the various component levels within <dsc>, and information is inherited from one hierarchical level to the next.

Namespace	http://www.tei-c.org/ns/1.0
Module	EAD
Attributes	att.EADGlobal (@id, @altrender, @audience, @encodinganalog) att.relatedencoding (@relatedencoding) att.desc.c (~~@level~~, @otherlevel)

@level The hierarchical level of the materials being described by the element.

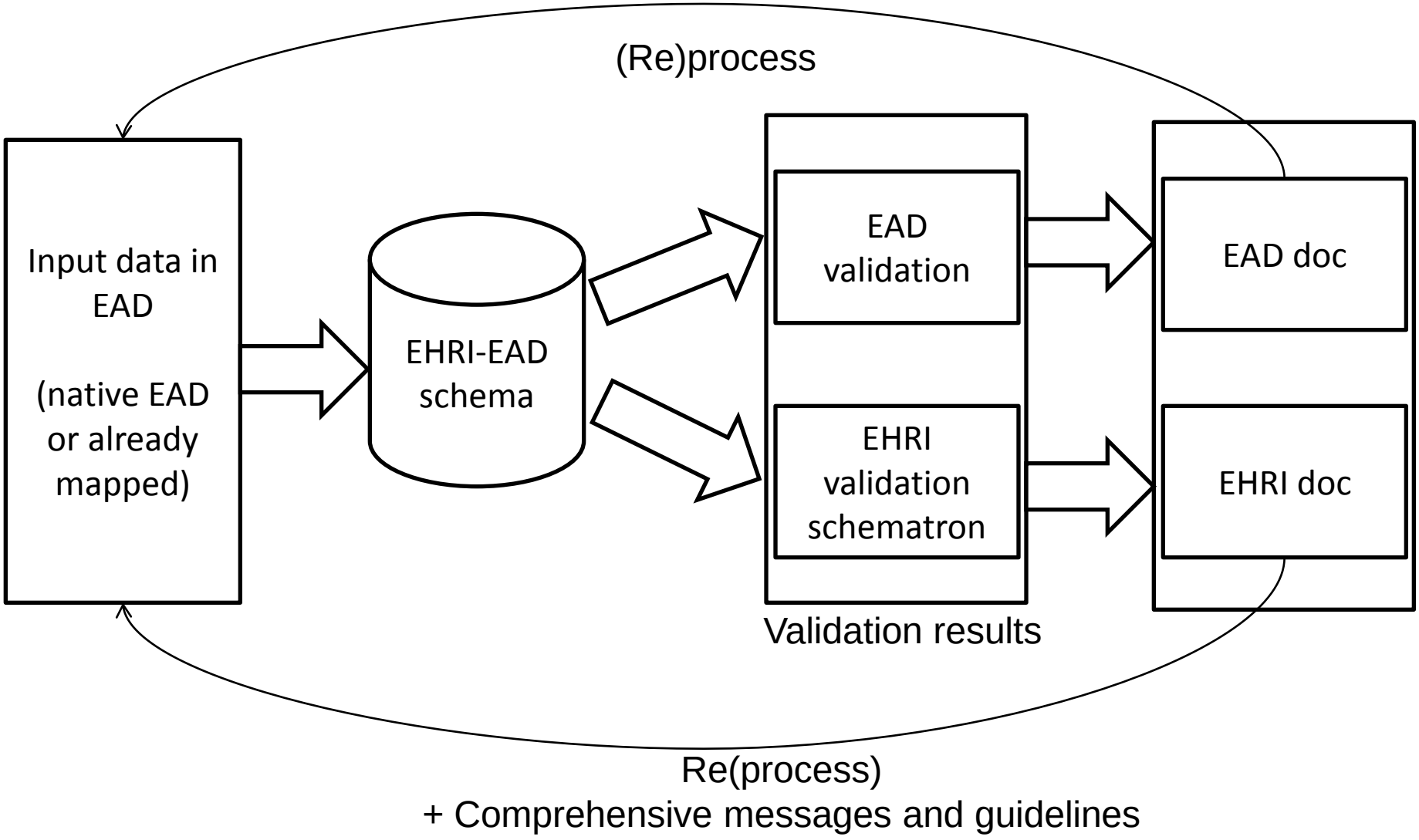
Derived from att.desc.c

Status **Required**

Schematron If the attribute @level has the value 'otherlevel', an attribute @otherlevel MUST be added

```
<s:rule context="ead:ead"
 see="https://cdn.rawgit.com/EHRI/data-validations/92c8e39f/ODD-RelaxNG/EAD/EHRI_EAD_doc.html#EAD.att.desc.c">
<s:assert role="MUST"
 test="not(@level = 'otherlevel') or (@otherlevel and not(@otherlevel = ''))">If the attribute level has the value 'otherle
 MUST be added</s:assert>
</s:rule>
```

Workflow


(Re)process
Input data in EAD
(native EAD or already mapped)
EHRI-EAD schema
EAD validation
EHRI validation schematron
Validation results
EAD doc
EHRI doc
Re(process)
+ Comprehensive messages and guidelines

Future developments

This method may be of a wider interest

1. Bridge between EAD2002 and EAD3 → soon an EAD3 ODD derived from EAD2002 ODD
2. Future maintenance of the EAD standard? in order to, like for the TEI, oriente this maintenance towards a (wise) continuous revision methodology.

Future developments

Conversion EAD → TEI

Several projects

First step: Mapping EAD <-> TEI (msDesc module)

→ Dominique Stutzmann, IRHT (Institut de recherches et d'histoires des textes) and Elena Pierrazzo, LUHCIE (Laboratoire universitaire Histoires Cultures Italie Europe)

More ambitious: extracting content from EAD with Regex, NER, … to create structured TEI descriptions beyond what EAD can provide.

→ Jean-Baptiste Camps (École des Chartes)

Remerciements à Marie-Françoise Limon-Bonnet, et toute l'équipe du projet Lectaurep (Convention Inria-Ministère de la culture)

EXTRACTION DE D'INFORMATION À PARTIR DES RÉPERTOIRES DES ÉTUDES NOTARIALES

LECTAUREP
Lecture Automatique de Répertoires

Archives nationales (DMC/DMOASI)
Équipe ALMAnaCH

Atelier Culture - Inria, 22 novembre 2018

Partenariat

Archives nationales

Département du Minutier Central (DMC)

Département de la maîtrise d'ouvrage du système d'information (DMOASI)

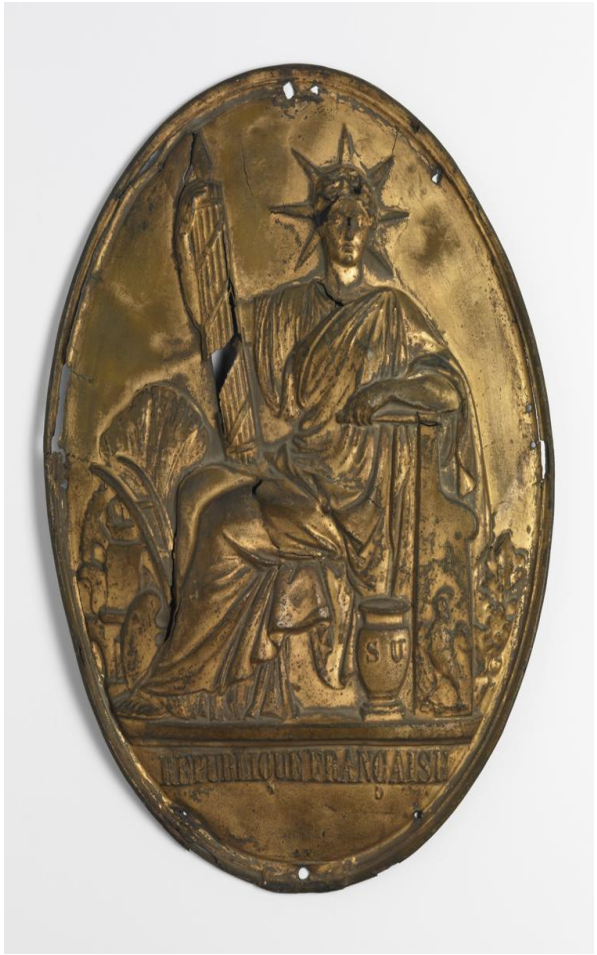
- Marie-Françoise Limon-Bonnet (DMC)
- Aurélia Rostaing (CC DMC)
- Danis Habib (CHED DMC)
- Frédéric Zamarreno (resp. DMOASI)
- Gaetano Piraino (DMOASI)

ALMAnaCH (*Automatic Language Modelling and Analysis and Computational Humanities*)

Équipe-projet Inria-EPHE

- Marie-Laurence Bonhomme, stagiaire Master TNAH de l'ENC
- Marc Bui, EPHE
- Éric de la Clergerie, Inria
- Benjamin Kiessling, EPHE
- Marie Puren, Inria
- Laurent Romary, Inria
- Charles Riondet, Inria
- Daniel Stökl Ben Ezra, EPHE

Le corpus d’archives



Le notaire, un officier public ministériel

- Détenteur d’une part de l’*auctoritas publica* ou de la foi publique,
- Producteur d’actes authentiques en forme de minutes,
- Lesquelles sont listées dans des répertoires

Le corpus d'archives

Les **répertoires des études notariales** sont des registres dans lesquels un clerc consigne, dans l'ordre chronologique et pour chaque jour, les actes qui ont été passés et enregistrés dans son étude.

Ils doivent nécessairement contenir la **date de l'acte**, sa **nature** et son **espèce** (en minute ou en brevet*) ainsi que le **nom des parties** et la **relation de l'enregistrement**.

Un formulaire pré-imprimé à 7 colonnes.

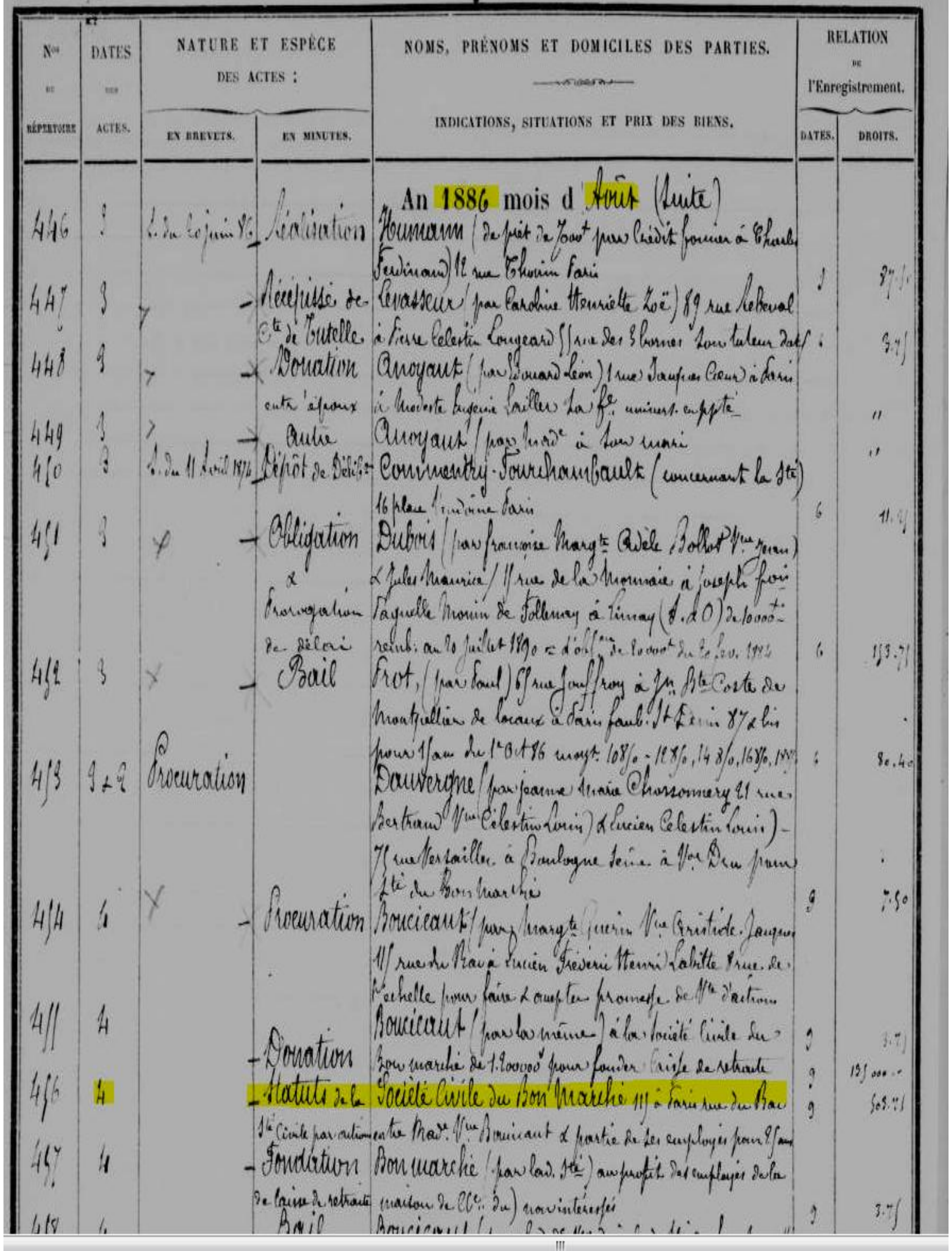
*La minute d'un acte notarié est l'original d'un acte authentique dont le notaire ne peut se défaire ; un acte en brevet est un acte notarié dont l'original est dépourvu de formule exécutoire, et est remis aux parties.

Un exemple de minute, les statuts de la société civile du « Bon Marché »

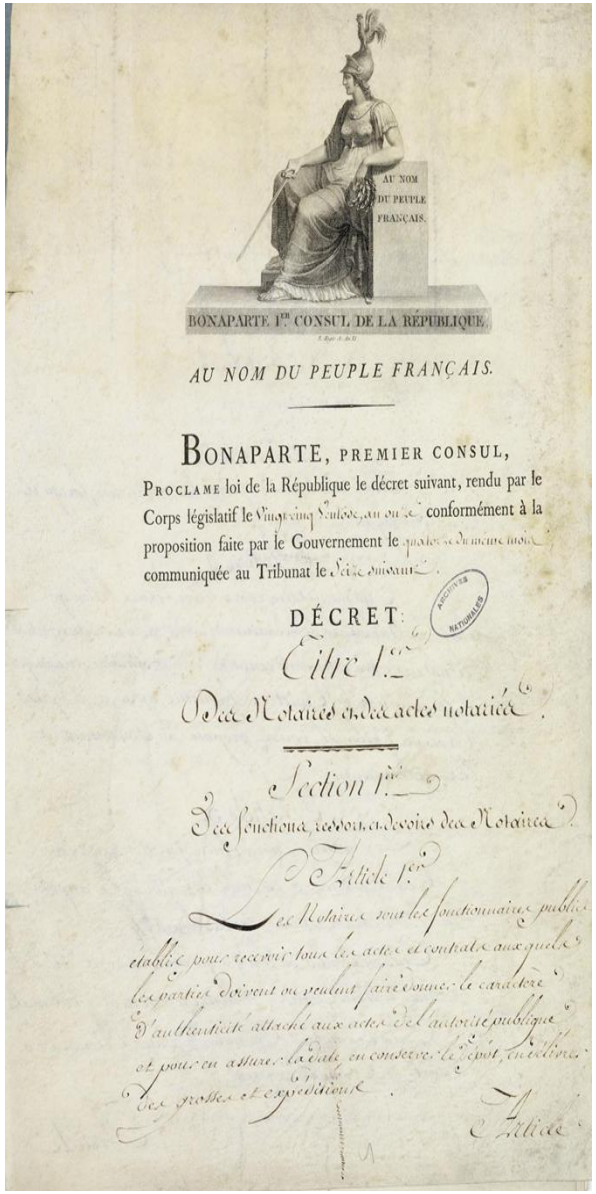
Arch. nat. MC/ET/XVI/1416

La mention de cette même minute - *Statuts de la société civile du Bon Marché* - au répertoire du notaire Gatine

https://www.siv.archives-nationales.culture.gouv.fr/siv/media/FRAN_IR_051822/c1ug18c3sei4--w32j38d5447t/DAFANCH96_016MIC06736_L



Le corpus d'archives



Depuis 1803 (25 ventôse an XI art. 29-30), le répertoire est « normalisé » et enrichi dans ses contenus

Projet LECTAUREP ⇒ Section du corpus

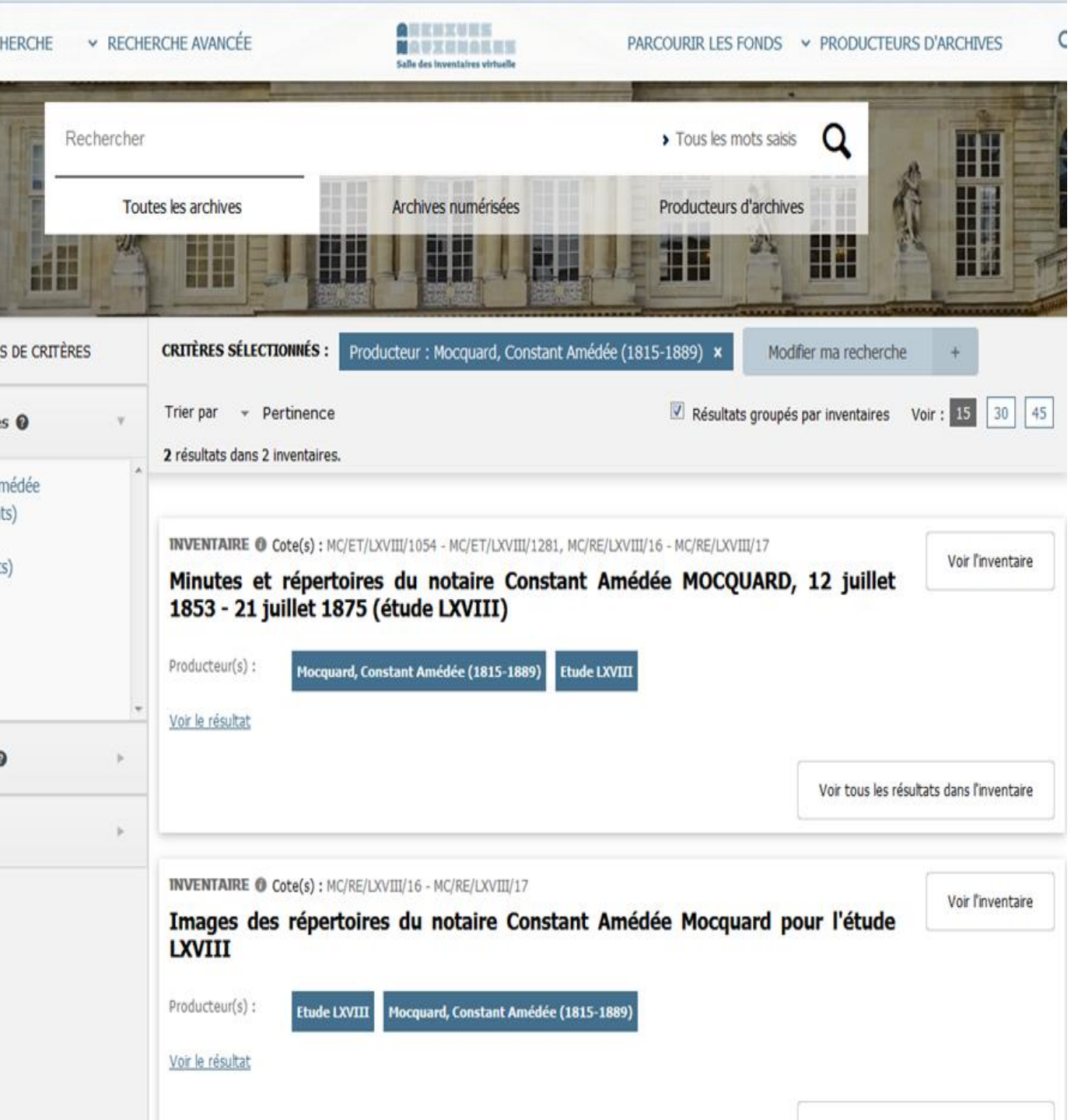
Entre 1803 et 1940, un corpus (rien qu'à Paris) de 917 notaires différents, env. 1800 registres

Le corpus d'archives

Le Minutier central des notaires de Paris, créé en 1928 au sein des Archives nationales. Il a pour vocation de conserver les fonds des 122 études notariales historiques de la capitale :

- 20 millions de minutes
- 3 300 répertoires
- 172 000 liasses et registres

⇒ 26 000 mètres linéaires.



Salle des inventaires virtuelle des Archives nationales :

Répertoires numérisées en ligne en mode image pour l'Ancien Régime comme pour le XIX^{e} et le premier XX^{e}.

N[os] du répertoire	Dates des actes	Nature et espèce des actes : En brevets	En minutes	Noms, prénoms et domiciles des parties — Indications, situations et prix des biens	Relation de l'Enregistrement : Dates	Droits
				An 1920, mois d'Avril		
804	13	Procuration		Brugnière (par Rose Suzanne Gaulhan, Vve de François Louis), à Paris rue Michel-Ange 84, — blanc, pour toucher	15	3.75
805	13		Donation	de Vaugrigneuse, (par [illegible] Charles Edmond) Vte [illegible], avenue Hoche 13, à Marie Anne Esther Tassel sa femme (toute ppté)	"	"
806	13		Donation	de Vaugrigneuse (par Mde) née Tassel, au survivant de son mari (toute ppté)	"	"
807	13		Ct de mariage	Lenormand (entre Henri Théophile Louis) à Paris rue Beauregard 26, et Marie Anne [illegible], Dt à Paris rue Montpensier 32 (Ct d'acquêts)	15	177 93
808	14	Procuration		Haberer (par Louis Eugène Albert), Dt à Paris rue Rambuteau 116, en blanc, pour radier inscript.	15	3.75
809	14	Ct de propté		Haas (carnet livret Caisse d'Epargne 75-321431 au nom de „ Loewenthal Lina", femme de Haas) au survivant, Dt à Vincennes rue du Bourg — de 250 f	16	"
810	14	mainlevée		Rodes (par Antoine) à Paris rue Gaillon 10 pré et actuelt rue St Lazare 13 d'inscript. au 2e de [illegible] [illegible] le 2 avril 1914 [illegible]		

Ressource riche mais difficile d'accès

Depuis le XVIe siècle, le répertoire est donc un registre qui liste, pour chaque jour, les actes passés dans une étude donnée...

… Pour exploiter ces répertoires, il faut toujours en passer par un dépouillement systématique parfois long

… Il n'en existe pas de transcription, pas de possibilité de recherche en plein texte

Objectifs du projet

- Offrir un meilleur accès à cette ressource

- Possibilité pour les chercheurs de réaliser des exploitations statistiques et quantitatives, en considérant ces répertoires non plus comme des inventaires mais comme des sources primaires.

Approche

- Reconnaissance automatique de structures et d'écriture manuscrite

- Mise à disposition des images numérisées de ces répertoires et de leur transcriptions sur une plateforme en ligne permettant d'y effectuer des recherches avancées.

Éléments de méthode

Segmentation des tableaux

- Distinction des classes d'information
- Analyse des répétitions de structures

HTR (Handwritten Text Recognition)

- Scribes multiples
- Utilisation de ressources sémantiques (référentiels des Archives Nationales)

Vers la conception d'une chaîne de traitement

1. Layout analysis : analyse de la structure physique et sémantique des répertoires
2. Reconnaissance d'écriture manuscrite
3. Extraction d'informations (Personnes, Lieux)
4. Brique participative (Correction des transcriptions, annotation)
5. Intégration dans le Système d'information des Archives Nationales
6. Mise à disposition de la communauté des chercheurs

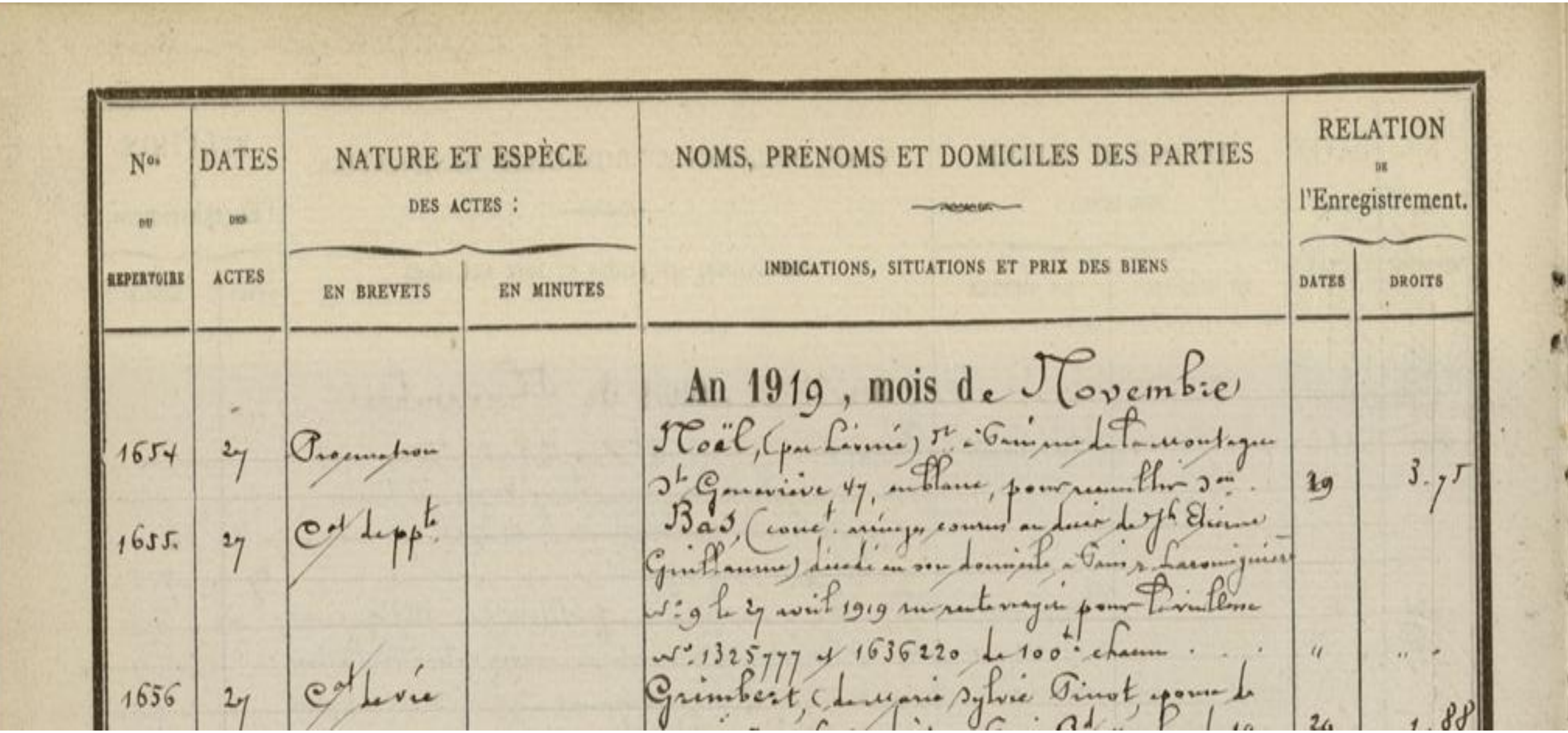
Phase 1 du projet

Accent mis sur les premières étapes de la chaîne de traitement

1. **Layout analysis : analyse de la structure physique et sémantique des répertoires**
2. **Reconnaissance d'écriture manuscrite**
3. *Extraction d'informations (Personnes, Lieux)*
4. *Brique participative (Correction des transcriptions, annotation)*
5. *Intégration dans le Système d'information des Archives Nationales*
6. *Mise à disposition de la communauté des chercheurs*

Layout analysis (structure de la page)

La segmentation des actes



Structure du tableau dans le registre :

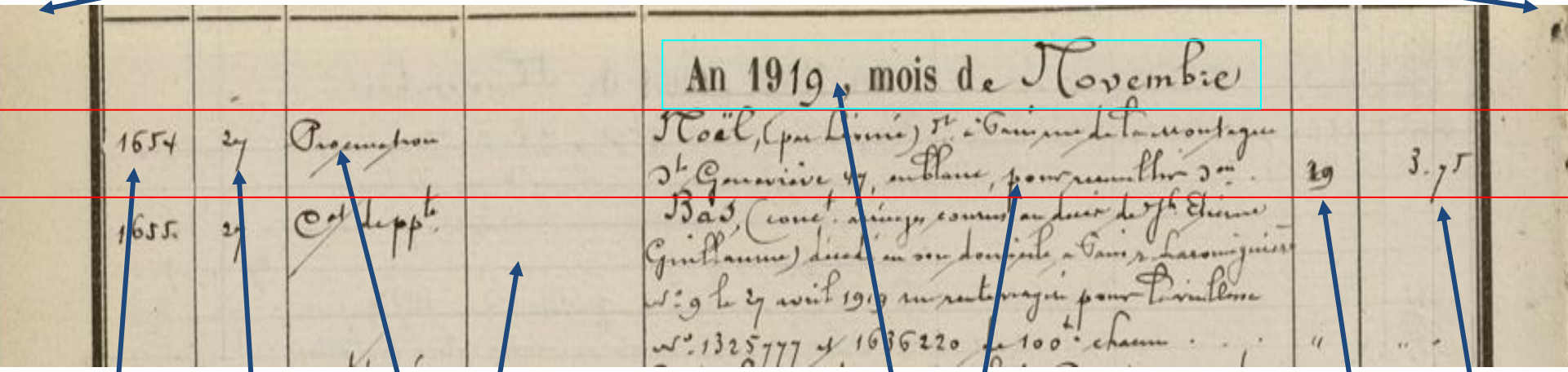
- Imprimé
- Lignes verticales
- Cohérent sur tout le corpus

Structure de l'acte :

- Manuscrit
- Plusieurs mains
- Pas de lignes horizontales
- Alignement aléatoire

La reconstitution des actes

Vide (sauf accident dans colonnes adjacentes)



N° de l'acte
Nombre entre 1 et 3000 (estimation)

Date de l'acte (jour)
Nombre entre 0 et 31

Typologie de l'acte
Chaîne de caractère
Vocabulaire contrôlé

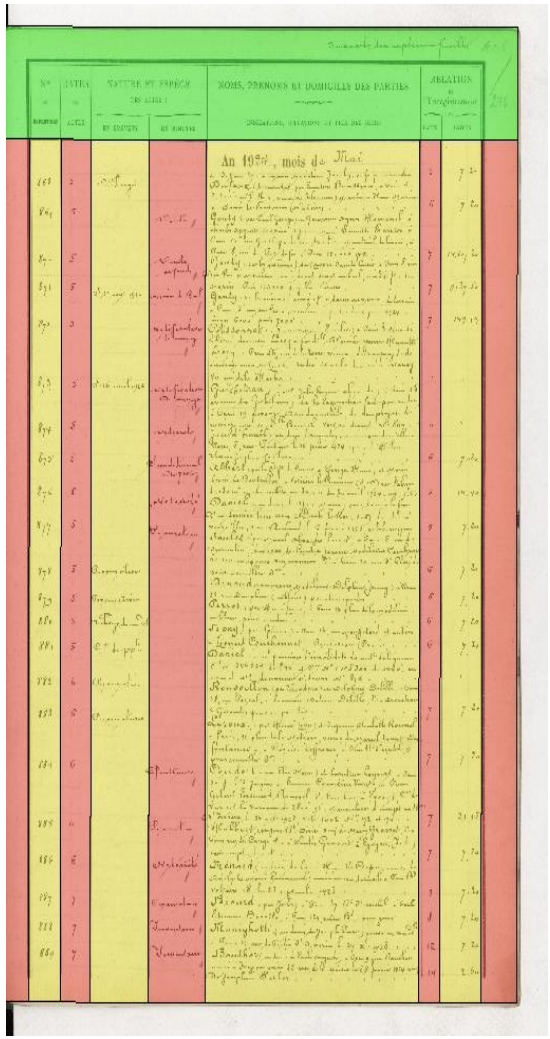
Date de l'acte (année et mois)
Écritures mixtes (imprimées et manuscrites

Description de l'acte
Nom et adresse des signataires, prix de vente d'un bien, date d'un décès, etc.

Date d'enregistrement (jour)
Nombre entre 0 et 31

Taxes acquittées
Chiffres, chaînes de caractères (gratis, etc.)

Segmentation automatique de la page, de l'entête et des colonnes

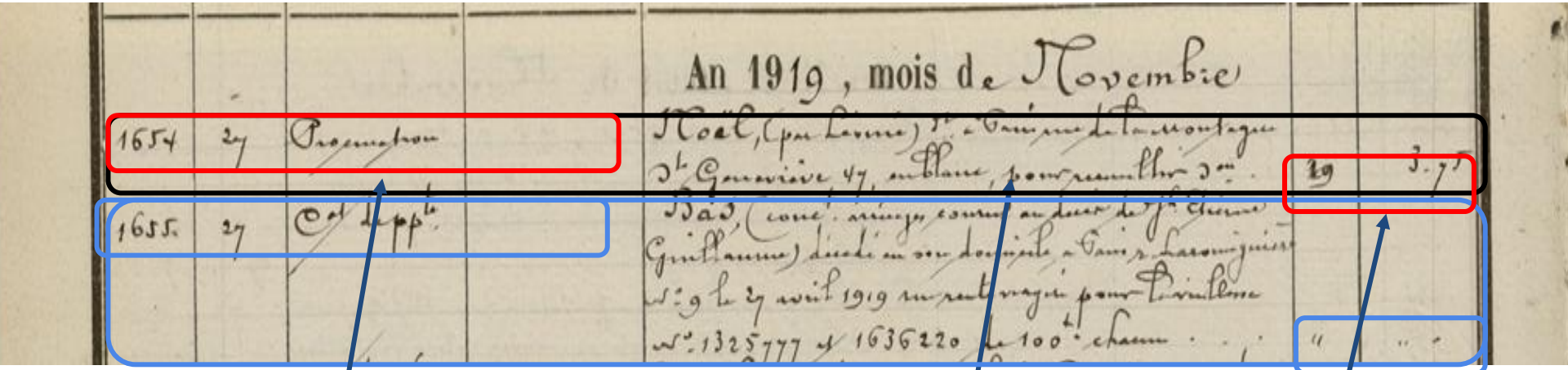


\+ tableau et images constants
\- la page peut-être en rotation

étapes

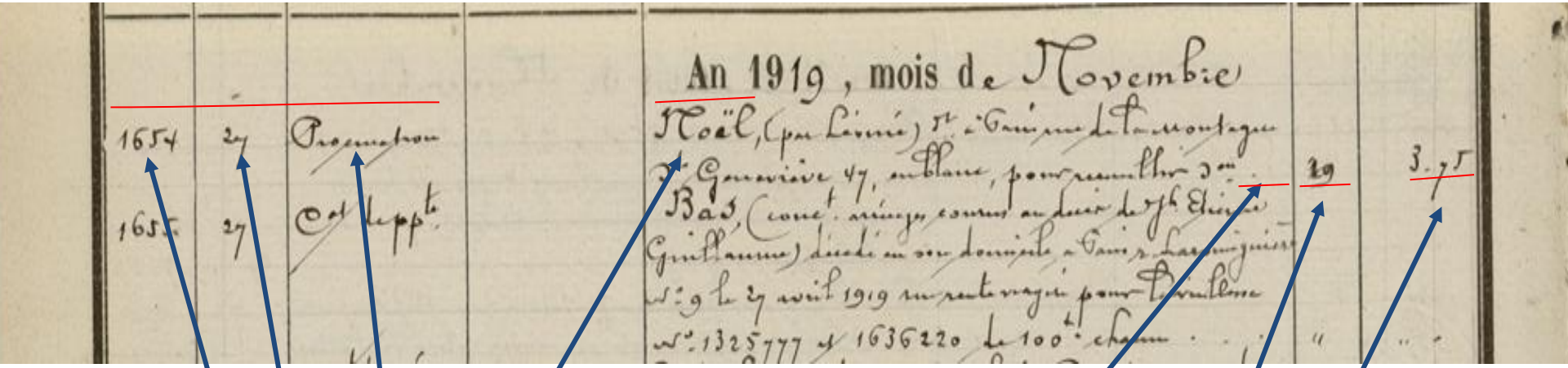
- détection de lignes horizontales et verticales du tableau (à l'aide de transformations morphologiques et de composantes connectés)
- détection et correction de la rotation de la page
- segmentation de la page de l'arrière-fond
- segmentation en colonnes et entête de la page
- suppression des séparateurs du tableau

Segmentation verticale des lignes et des actes



Début de l'acte	Colonne centrale : Description de l'acte	Fin de l'acte
écriture à la première ligne de l'acte	*zone avec écriture intensive. lignes remplies*	*écriture à la dernière ligne de l'acte*

Indices de segmentation



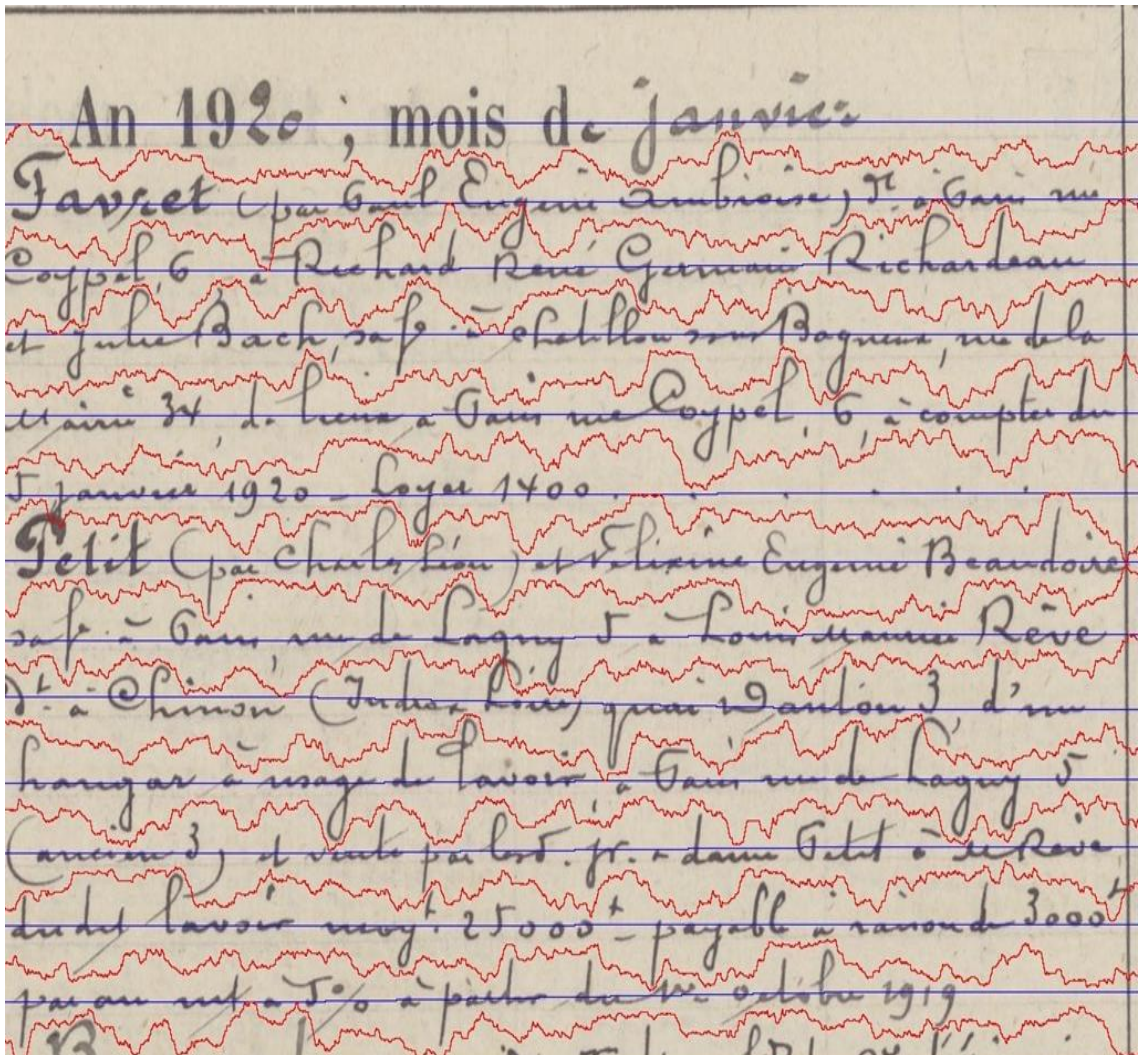
Limite supérieure de l'acte:

- *N° d'ordre*
- *Date de l'acte*
- *Typologie*
- *Nom du signataire en gras*

Limite inférieure de l'acte:

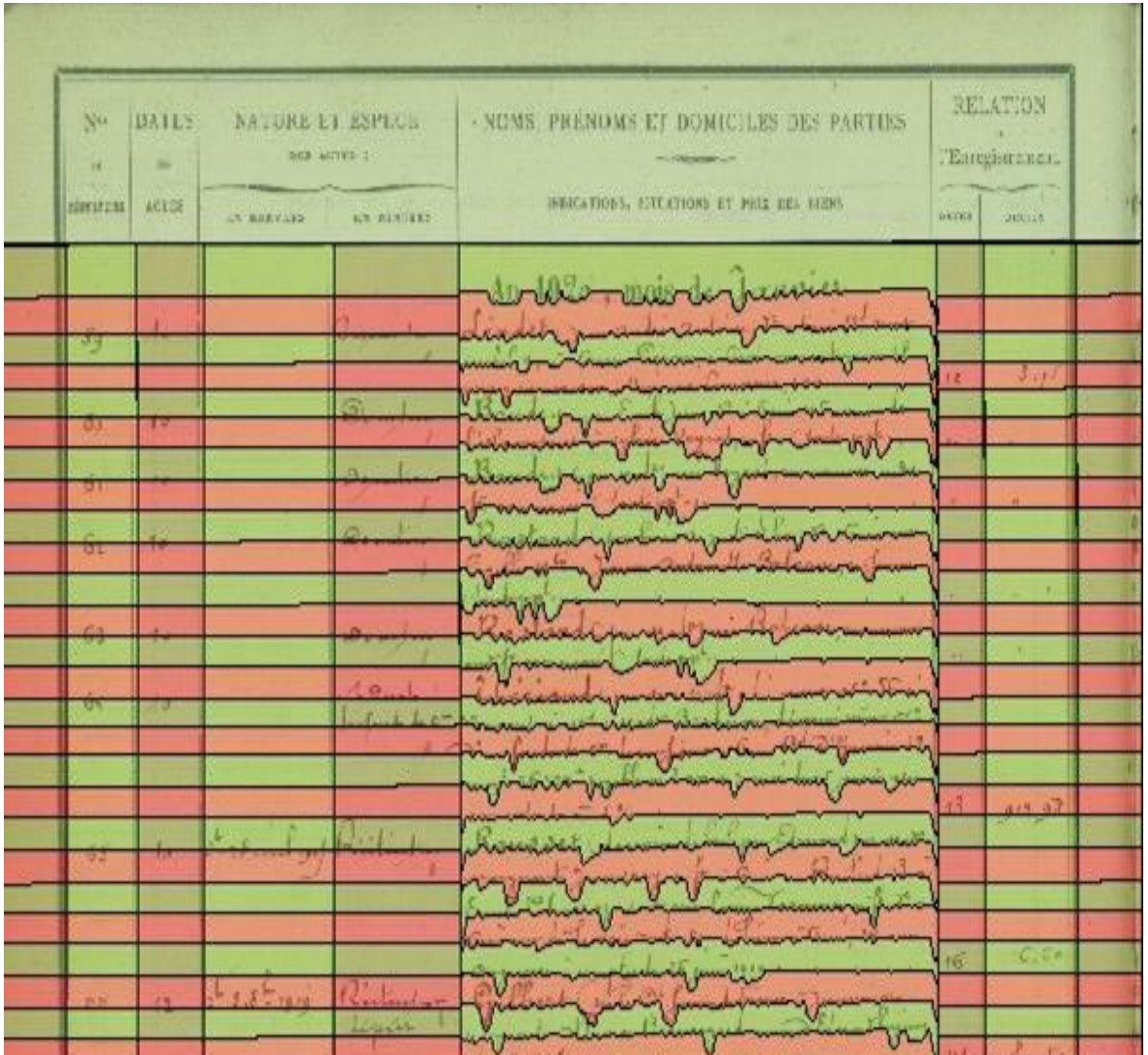
- *Description de l'acte: dernière ligne parfois incomplète*
- *Date de l'enregistrement de l'acte*
- *Taxes acquittées*

Segmentation automatique des lignes



Seamcarving (chemin d'énergie minimale) pour segmenter les lignes de la colonne de description de l'acte.

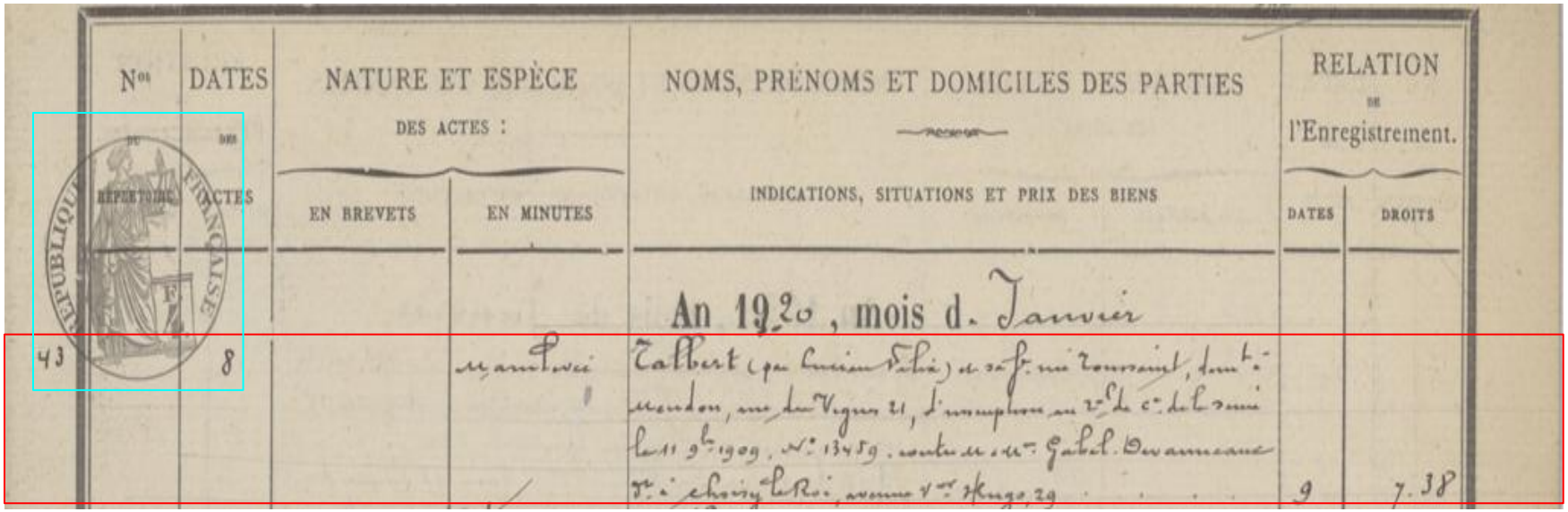
Extension de la segmentation des lignes au reste du tableau



Bons résultats en général:
Alignements horizontaux parfois décalés

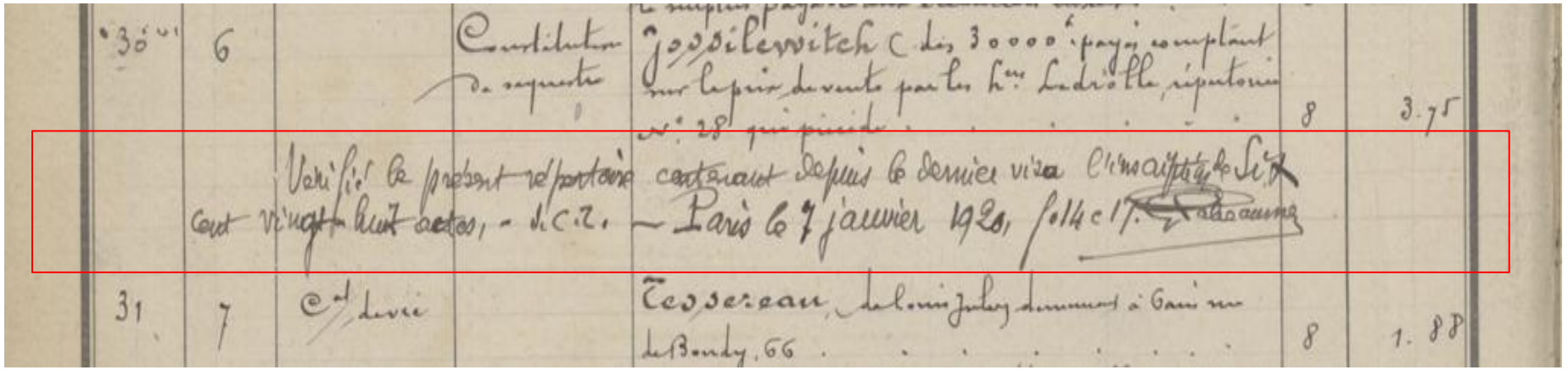
Difficultés de segmentation

Présence d'un obstacle (tampon) dans la colonne 2 ⇒
Numéro de l'acte décalé dans la marge (colonne 1)



Difficultés de segmentation

- Dépassement du texte sur plusieurs colonnes
- Ruptures textuelles aussi lors des mentions introductives ou légales du juge



Reconnaissance d'écriture manuscrite

HTR et Classification

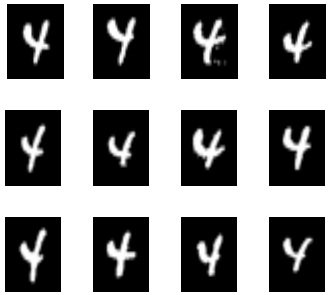
1. Classification par reconnaissance de formes
   - Colonnes de chiffres ou de nombres
     - Dates
     - N° d'actes
     - Taxes
   - Colonnes de texte « simple » des types d'actes

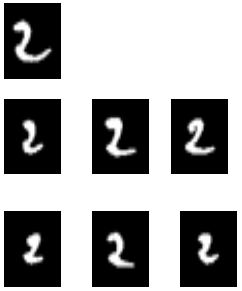
2. HTR : Obtention de données d'entraînement pour construire un modèle de reconnaissance
   - Transcription collaborative d'une cinquantaine de pages avec *Transkribus* d'un répertoire de l'étude Marotte (importation des images numérisées, segmentation, transcription cellule par cellule)

Classification

Classifications par reconnaissance de formes

- Chiffres (notamment numéros de jour) : 98,66 % de précision (pour les chiffres de 0 à 9) grâce à un algorithme de classification basé un réseau de neurones convolutif





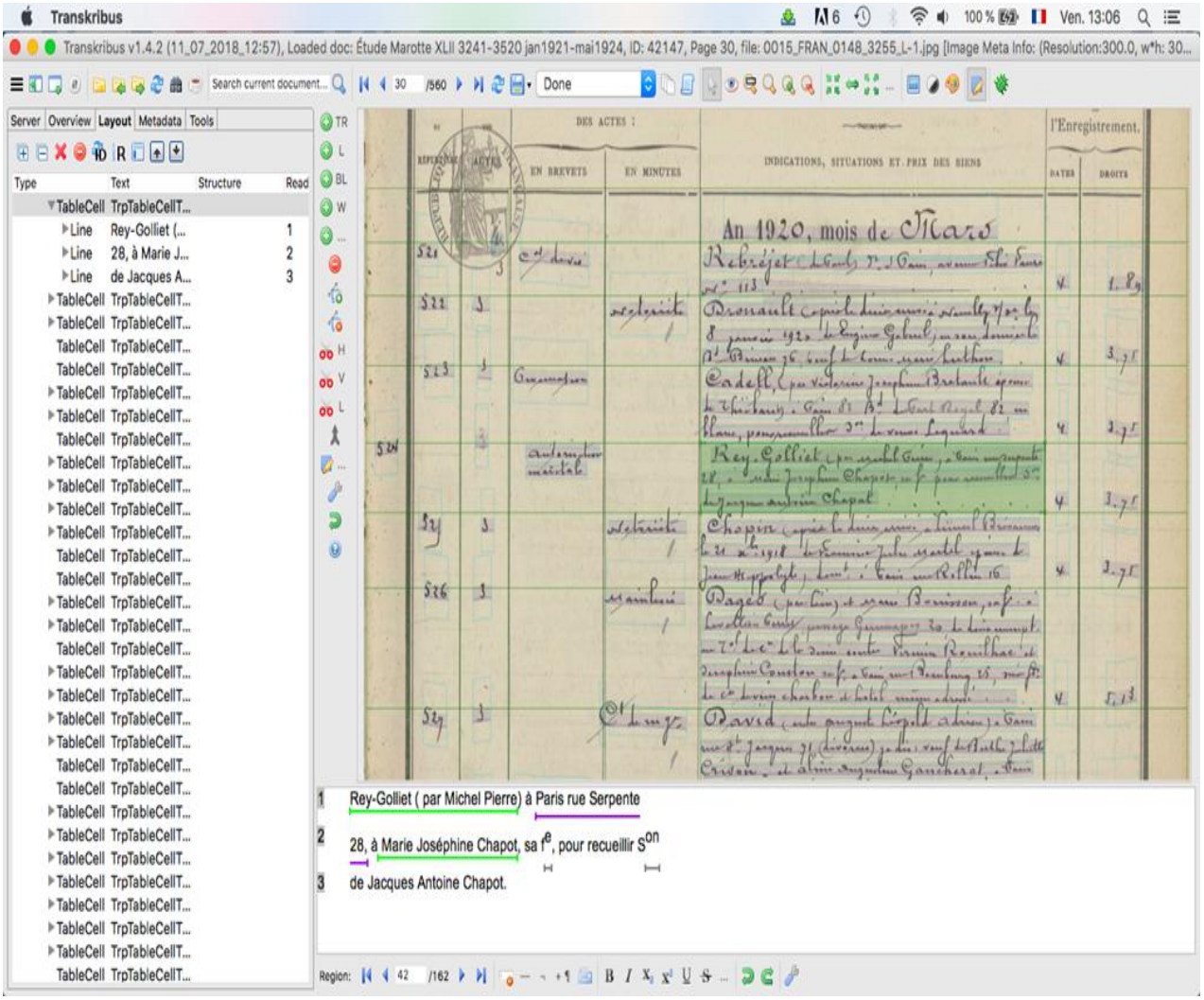
Classification

Typologies d'actes : Premier travail de classification manuelle à poursuivre. Difficultés :

- Hétérogénéité des formes
- Typologies multiples pour un acte
- Typologie absente (quand un acte est la "suite" d'un premier acte)

Problèmes	**Forme "normalisée"**	**Variation(s)**	
Formes hétérogènes	Contrat conditionnel	*Ct conditionel*	*Contt condel*
	Contrat de mariage	*Ct de mariage*	*Ct de mge*
Types multiples	Liquidation	*Liquidation et partage*	*Liquidation et présentation de compte de tutelle*
Acte faisant suite à un 1er	Suite de l'acte du 15 décembre 1919	*Ste 15 xbr 1919*	

HTR: Acquisition des données d'entraînements (transcription et annotation)



Transkribus.eu

- \+ Facile d'utilisation
- \+ Travail collaboratif
- \+ Formats d'export variés
- \- Customisation faible
- \- Pas de maîtrise du stockage des données et des modèles entraînés
- \- Modèles non open-source
- \- Futur incertain

HTR: Entraînement et premiers résultats

- Entraînement d'un modèle HTR à partir des pages transcrites manuellement "*vérité terrain*" (un seul scribe) :
  - M1 : 40 pages
  - M2 : 50 pages (1 million de mots)
- Character Error Rate (CER) sur un échantillon test du même répertoire:
  - M1 : **13,5 %**
  - M2 : **10,4 %**

  ⇒ ce qui reste considérable, mais peut être amélioré
- Le modèle M2 a été testé sur quelques pages d'autres registres et donc d'autres mains. Les résultats ne sont naturellement pas bons (CER autour de 40 %) : nécessité de données d'entraînement plus hétérogènes.

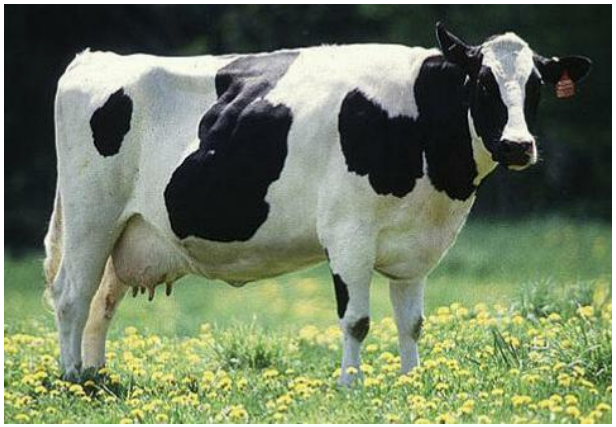
Défis de la phase 2

- Acquisition d'autres données d'entraînement pour d'autres mains, d'autres périodes et créer un modèle mixte
- Passage en production
  - ⇒ Envisager une plateforme de transcription collaborative
  - ⇒ Collaboration avec le projet SCRIPTA-PSL (transcription automatique de manuscrits historiques multilingues)
  - ⇒ Utilisation du logiciel d'HTR Kraken (Benjamin Kiessling, ALMAnaCH) pour plus de contrôle et d'autonomie
- Utilisation des données sémantiques
  - ⇒ Référentiels spécifiques au corpus : typologie des actes, terminologie notariale
  - ⇒ Référentiels d'autorités pour les entités nommées (personnes et lieux : utilisation d'annuaires)
- Expérimentation sur les répétitions (word spotting)

Remerciements à Patrice Lopez, Luca Foppiano

STRUCTURATION AUTOMATIQUE DE CONTENUS

Aller au delà des données primaires non structurées (PDF)



Cow (structured data)

publishing

text-mining

Hamburger
(unstructured data)

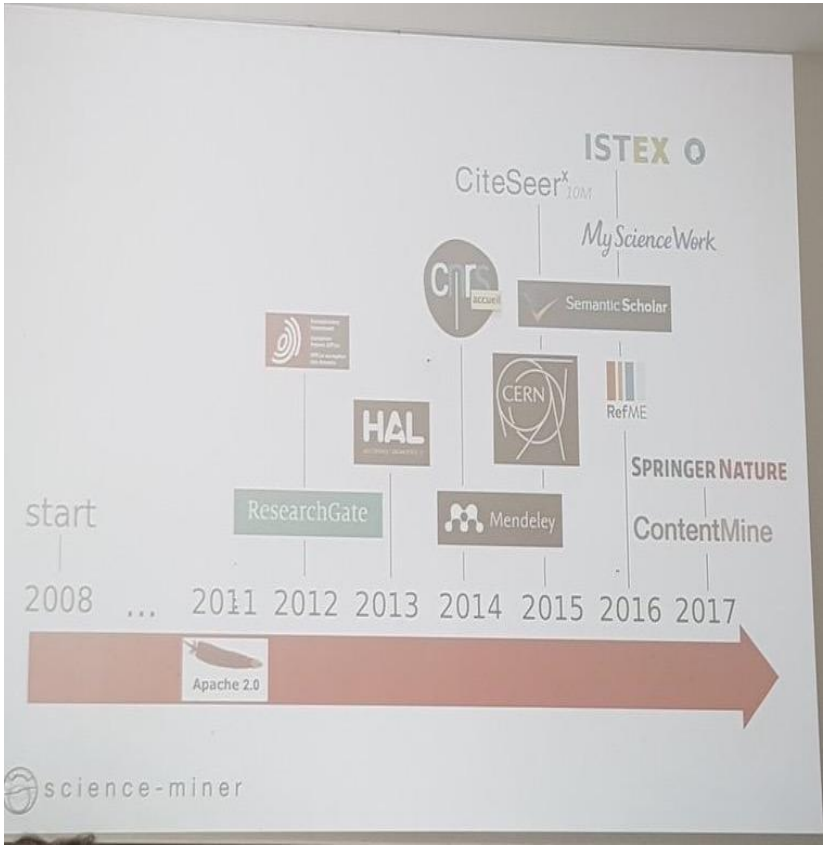
"Converting PDF to XML is a bit like converting hamburgers into cows. You may be best off printing it and then scanning the result through a decent OCR package."
Michael Kay (http://lists.xml.org/archives/xml-dev/200607/msg00509.html)

Inspired from: Duncan Hull

Getting acquainted to GROBID

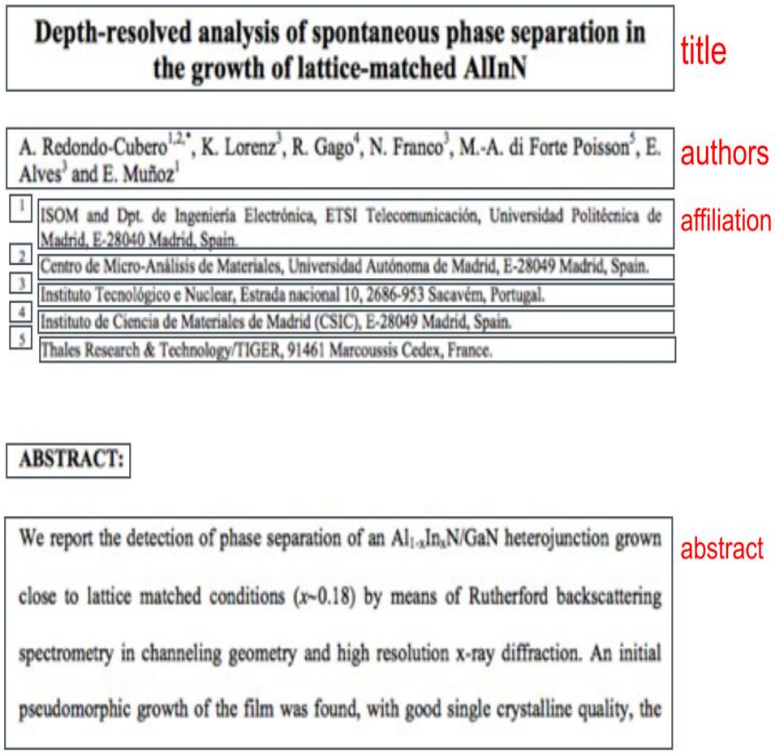
GROBID (GeneRation Of Bibliographic Data) (*Lopez et al. 2015*)

- Cascading content extraction from PDF
- CRF: Conditional Random Fields
- TEI: corpus annotation and final output



Example: GROBID for meta-data extraction

- GROBID (GeneRation Of Bibliographic Data) (*Lopez et al. 2015*)



Depth-resolved analysis of spontaneous phase separation in the growth of lattice-matched AlInN

title

A. Redondo-Cubero[1,2,*], K. Lorenz[3], R. Gago[4], N. Franco[3], M.-A. di Forte Poisson[5], E. Alves[3] and E. Muñoz[1]

authors

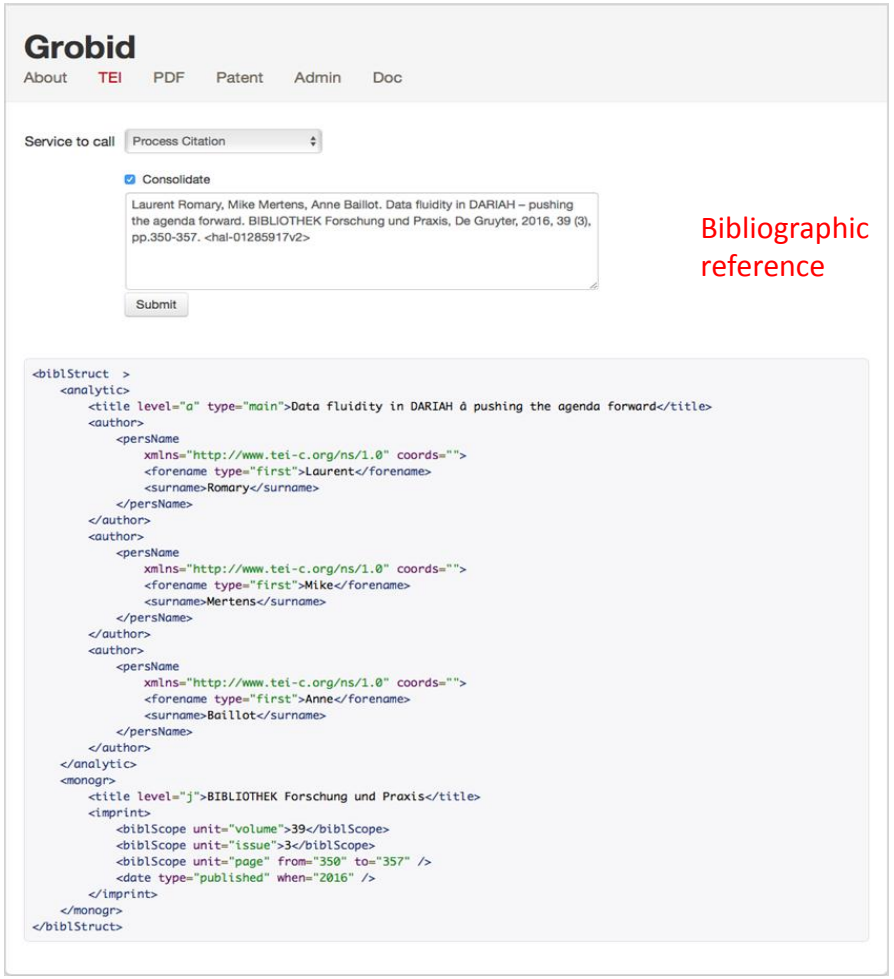
1 ISOM and Dpt. de Ingeniería Electrónica, ETSI Telecomunicación, Universidad Politécnica de Madrid, E-28040 Madrid, Spain.
2 Centro de Micro-Análisis de Materiales, Universidad Autónoma de Madrid, E-28049 Madrid, Spain.
3 Instituto Tecnológico e Nuclear, Estrada nacional 10, 2686-953 Sacavém, Portugal.
4 Instituto de Ciencia de Materiales de Madrid (CSIC), E-28049 Madrid, Spain.
5 Thales Research & Technology/TIGER, 91461 Marcoussis Cedex, France.

affiliation

ABSTRACT:

We report the detection of phase separation of an $Al_{1-x}In_xN$/GaN heterojunction grown close to lattice matched conditions (x~0.18) by means of Rutherford backscattering spectrometry in channeling geometry and high resolution x-ray diffraction. An initial pseudomorphic growth of the film was found, with good single crystalline quality, the

abstract



From GROBID to GROBID-Dict

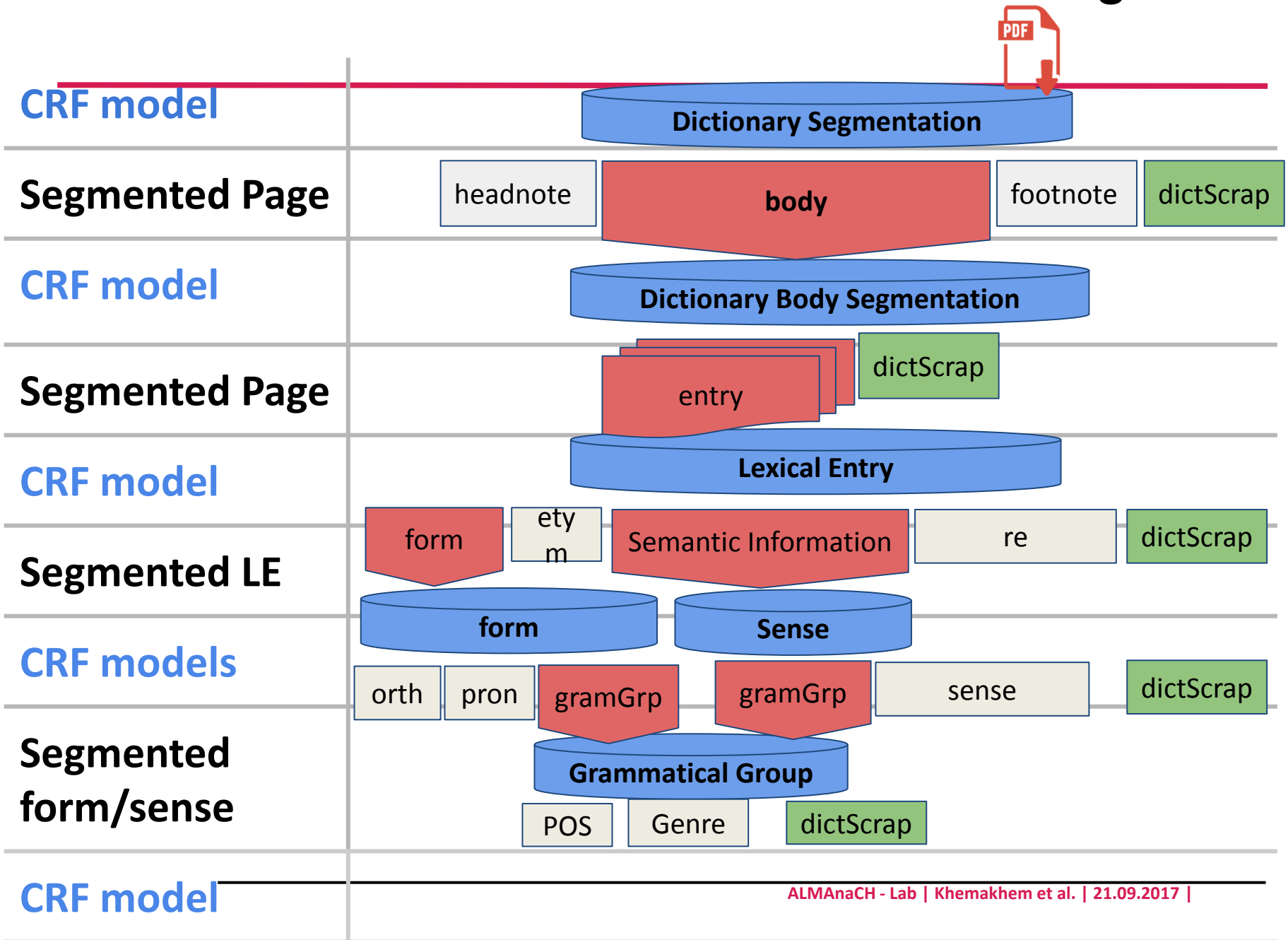
- Numerous projects dealing with legacy (unstructured) dictionaries
  - Monolingual, Bilingual
  - Old, modern
  - Cf. Borchmann et alii, Widmann & Buchanan
- Possible transition?
  - Costly manual and rule based techniques
  - Machine Learning (ML)
- Need for exchangeable lexical resources (LR)s
  - TEI…

Approach

- Cascading extraction models

- Using a univocal TEI format
- Taking into account the standardisation context:
  - Towards a terser TEI subset
  - TEI-Lex 0, ISO 24613-1
  - Highlights:
    - No (less) <entryFree>, <superEntry>
    - Reduced number of elements to encode a Lexical Entry (LE)
    - ...

GROBID-Dictionaries: LI Processing



Orkney Scots dictionnary

chap *v.* **1** knock, '*He chappid fower or five times at the door and got no reply.*' **2** mash potatoes, '*I like me tatties chappid.*' **3** chop wood, '*Ah'll go and chap twa three sticks for the fire.*' **chappeen tree** potato masher.

```
<entry>
  <form type="lemma">
    <orth>chap</orth>
  </form>
  <gramGrp><pos>v.</pos></gramGrp>
  <sense>
    <sense>1 knock, 'He chappid fower or five times at the door and got no reply.'</sense>
    <sense>2 mash potatoes, 'I like me tatties chappid.'</sense>
    <sense>3 chop wood, 'Ah'll go and chap twa three sticks for the fire.' chappeen tree
potato masher</sense>
  </sense>
</entry>
```

Mueller (1878) (“Etymologisches Wörterbuch der englischen Sprache”, 2nd ed., Cöthen 1878/1879)

Cabbage 1. *kohl; altengl.* cabage, *bei* Hal. 226 cabes, cabishes: *mlat.* gabusia, *fr.* cabus, *it.* cappuccio; *vgl. ndl.* cabuis, cabuyscoole, *nhd.* kappes, *worüber* Weigand 1, 562: „*Im* vocab. incip. teut. ante lat. kabbas, *mhd.* der kapaʒ, kapeʒ, *spätahd.* kabuʒ, capuʒ. *Aus fr. der* cabus, *it.* capúccio, *welches wie russ. die* kapusta *kohl, aus mlat.* caputium *kapuze hervorging und der geschlossene kohl schien einer mönchskappe ähnlich;“ vgl.* Diez 1, 110 *und unter den nhd.* kabisz, kabis Grimm 5, 9.

```
<entry>
   <form>
      <orth>Cabbage</orth><label>1.</label>
   </form>
   <etym>
      <seg><def>kohl</def>;<lang>altengl.</lang>
         <mentioned>cabage</mentioned>,<seg>bei</seg>
         <bibl>Hal. 226</bibl><mentioned>cabes</mentioned>,
         <mentioned>cabishes</mentioned>: <lang>mlat.</lang>
         <mentioned>gabusia</mentioned>, <lang>fr.</lang>
         <mentioned>cabus</mentioned>, <lang>it.</lang>
         <mentioned>cappuccio</mentioned>;
         <seg>vgl.</seg><lang>ndl.</lang>
         <mentioned>cabuis</mentioned>,
         <mentioned>cabuyscoole</mentioned>, <lang>nhd.</lang>
         <mentioned>kappes</mentioned>, <seg>worüber</seg>
         <bibl>Weigand 1, 562</bibl>:
      </seg>
      <quote>„Im <bibl>vocab. incip. teut.</bibl> ante
         <lang>lat.</lang> <mentioned>kabbas</mentioned>,
         <lang>mhd.</lang> der <mentioned>kapaʒ</mentioned>,
         <mentioned>kapeʒ</mentioned>,<lang>spätahd.</lang>
         <mentioned>kabuʒ</mentioned>,
         <mentioned>capuʒ</mentioned>.Aus <lang>fr.</lang>
         der <mentioned>cabus</mentioned>, <lang>it.</lang>
         <mentioned>capúccio</mentioned>, welches wie
         <lang>russ.</lang> die <mentioned>kapusta</mentioned>
         <def>kohl</def>, aus <lang>mlat.</lang>
         <mentioned>caputium</mentioned> <def>kapuze</def>
         hervorging und der geschlossene kohl schien einer
         mönchskappe ähnlich;“
      </quote>
      <seg><seg>vgl.</seg> <bibl>Diez 1, 10</bibl>
         <seg>und unter den</seg> <lang>nhd.</lang>
         <mentioned>kabisz</mentioned>,
         <mentioned>kabis</mentioned> <bibl>Grimm 5, 9</bibl>.
      </seg>
   </etym>
</entry>
```

Projet Dopamine

EPILOGUE: TRACER LES DONNÉES D’EXPÉRIENCE

Un contexte national et européen « pressant »

- Le Plan national pour la science ouverte – juillet 2018
  - « Rendre obligatoire la diffusion ouverte des données de recherche issues de programmes financés par appels à projets sur fonds publics »
  - « La France recommandera l'adoption de licences ouvertes pour les publications et les données »
- De nombreuses initiatives européenne
  - Publications ouvertes : OpenAire, Plan S
  - Données
    - Obligation de la production d'un plan de gestion de données, pression pour disposer de données « FAIR »
    - RDA, GO FAIR, mise en place d'EOSC
  - Infrastructures européennes de la feuille de route ESFRI: DARIAH, E-RIHS, CLARIN, OPERAS, DiSSCo

Expérience acquise dans le cadre d'Iperion CH

- Iperion CH
  - Integrated Platform for the European Research Infrastructure ON Cultural Heritage
  - INFRAIA-1-2014-2015 - Integrating and opening existing national and regional research infrastructures of European interest
  - 19 équipements, dans 11 pays regroupés au sein de 3 plates-formes: ARCHLAB, FIXLAB and MOLAB
- Enquêtes effectuées au sein de la tâche 2.2 *Management plan of generated digital data*
  - Collaboration avec IPANEMA

Enquête Iperion CH – principaux résultats

- Double enquête sur les pratiques et les jeux de données
  - excellente couverture du consortium
  - 3 plates-formes et 29 instruments, 78 jeux de données mentionnées couvrant une très grande variété de matériaux et de méthodes d'analyse
- Difficultés exprimées par les répondants
  - documentation des données
  - formats standards et réutilisables
  - hébergement informatique pérenne
  - licences associées aux jeux de données

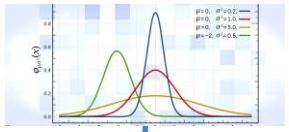
Et le chercheur/la chercheuse dans tout ça ?

- Recentrer le débat sur l'individu et le projet de recherche
  - Les principes FAIR sont trop centrés sur les jeux de données, mais pas sur les acteurs du processus de recherche
  - L'administration des données ne doit pas prendre le pas sur la recherche elle-même
- Identifier les questions qui peuvent se poser sur le terrain
  - Documentation, attribution, hébergement, réutilisation
- Apporter des réponses simples et concrètes
  - Intégrer la gestion des données à la pratique de recherche

Accompagner le processus de recherche – vers un *certificat d'identité des données*

- Vision: agréger à chaque étape tous les éléments pertinents pour la traçabilité d'un jeu de données
  - Acteurs et responsabilités
  - Sources
  - Données techniques, formats
  - Traitements effectués
  - Conditions de diffusion et de réutilisation, licences
- Démarche pragmatique
  - Identifier comment rendre cette gestion naturelle (mais pas transparente) pour la chercheuse ou le chercheur

Tracer la création des jeux de données



Projet de recherche

Jeux de données primaires

Jeux de données secondaires

Publication

Echantillon
Origine
Propriétés

Appareil
Format
Paramètres utilisés

Logiciel
Format
Paramètres utilisés

Littérature secondaire

Littérature primaire

Documentation

Données d'enquête

Et quand on croit être arrivé au bout…

- Hébergement
  - Données, méta-données
  - Identification, archivage à long terme
  - Autorité?
- Référencement (publications)
  - Citer les jeux de données (et sources) primaires et secondaires
  - Citer les différents acteurs du processus
  - Réutiliser les contenus?

Aborder le système en amont

- De la demande initiale à la réalisation du projet de recherche
- Possible utilisation de SciencesCall
  - Une plate-forme de gestion d'appel à projet
    - Développement à l'UMS CCSD (cf. HAL, Episciences, SciencesConf)
    - Construit suivant un modèle classique de dépôt et évaluation

Pistes pour une utilisation de SciencesCall pour le traçage des données

- Fiche de gestion de données au moment du dépôt
- Intégration d'un point de validation par les acteurs du processus
  - Etablissement patrimonial, équipement, futur hébergeur des données
- Agrégation de métadonnées de la part de ces acteurs

Intégrer tous les acteurs du dispositif

- Le chercheur, la chercheuse
  - Qualifie la recherche et définit le cycle de vie de ses données
- L'établissement patrimonial
  - Détermine les contraintes liées à l'utilisation de ses fonds et les attentes en retour
- L'équipement
  - Détermine une politique d'utilisation de la part des projets de recherche
- L'hébergeur de données
  - Exprime des contraintes sur la taille, les formats, la documentation, les conditions d'accès
- Et bien sûr les politiques européennes, nationales et institutionnelles

Un outil méthodologique

- La charte de réutilisation des données
  - un contrat entre les différents acteurs du processus de recherche
  - fluidifier les échanges et faciliter la réutilisation des données patrimoniales
- Implication d'institutions européennes
  - DARIAH, Europeana, CLARIN, APE, E-RIHS

Les principes fondateurs

Tentative d'implémentation

- Partir des rôles et des exigences variés entre acteurs pour faire des recommandations centrées sur les besoins de chacun, mais profitables à tous (approche *bottom-up*)
- Par exemple, pour la « *citability* » :
  - Pour le chercheur ou la chercheuse → citer l'institution d'où viennent les données et l'équipement / l'infrastructure qui est intervenue au cours du projet
  - Pour l'institution patrimoniale → fournir un modèle de citation et communiquer autour des recherches faites à partir de leurs collections
- Tester ces principes à travers **SciencesCall**, plateforme de gestion des appels à projet

Archives départementales des Yvelines

Conditions de réutilisation des documents conservés aux Archives départementales des Yvelines

Vous êtes libre de réutiliser gratuitement et sans formalités les informations contenues dans les documents conservés aux Archives départementales ou les images de ces documents,

Vous pouvez les :

- reproduire, copier, publier et transmettre ;
- diffuser et redistribuer ;
- adapter, modifier, transformer, notamment pour créer des documents dérivés ;
- exploiter à titre commercial.

La réutilisation est gratuite mais la mise à disposition des informations donne lieu à la perception de frais techniques dans les cas où elle entraîne des opérations techniques (reproduction, extraction de données, compression et transfert de fichiers, …) à la charge du département.

Source: https://archives.yvelines.fr/article.php?larub=12&titre=reutilisation-des-archives

Archives Départementales des Yvelines (Suite)

Sauf s'il s'agit de documents dont vous avez obtenu communication par dérogation, de documents contenant des informations publiques comportant des données à caractère personnel, de documents sur lesquels s'exerce un droit de propriété intellectuelle, ou encore de documents entrés par don ou par dépôt.

…

CAS DES DOCUMENTS COMPRENANT DES DONNÉES À CARACTÈRE PERSONNEL

…

CAS DES DOCUMENTS SUR LESQUELS S'EXERCE UN DROIT DE PROPRIÉTÉ INTELLECTUELLE

…

Et à condition de respecter les conditions suivantes :

- ne pas dénaturer le sens des informations contenues dans les documents ;
- mentionner de manière visible la source des informations et leur lieu de conservation (de préférence sous la forme Archives départementales des Yvelines, précision de la cote) ;
- préciser la date de la production ou de la dernière mise à jour des informations ;
- mentionner le nom de (ou des) auteur(s), s'il y a lieu.

Et pratiquement?

- Priorité au travail avec les chercheurs
  - Doctorants et post-doctorants financés dans le cadre du DIM MAP
    - Séminaire de travail commun le 7 février
    - Echanges personnalisés en mai-juin
- Travailler à une déclaration préalable à intégrer aux demandes de projet sur SciencesCall
  - Intégration des grandes catégories de la charte
- Définir un concept pour intégrer des contraintes multi-acteurs dans ScienceCall
- Vers un guide de gestion des données de recherche à destination des jeunes chercheurs
  - Centré sur les processus de recherche et les types de donnée

Perspectives

- Gestion raisonnée des données
  - Un point de départ pour la transparence scientifique
    - Dans l'espace et le temps
  - Nécessite des compétences et des infrastructures
  - Du temps, de l'argent, des personnels
- Aborder la gestion des données de façon humaine et pragmatique
  - Accompagner le changement en impliquant les chercheurs
  - Permettre une amélioration progressive des conditions de gestion des données de la recherche
  - Mettre en œuvre des solutions pérennes