Du document physique à la base de données : les chaînes de numérisation, procédés et problématiques.

Morgane Pica

Du document physique à la base de données :

les chaînes de numérisation, procédés et problématiques.

RÉGION NORMANDIE

CONDE

CRISCO

Humanités numériques

- Discipline au croisement de la recherche en Sciences Humaines et Sociales et des techniques numériques.
- Formation en général hybride : formation théorique dans la recherche et formation pratique en techniques du numérique.

Humanités numériques

« Les HN peuvent s'appuyer sur une histoire passée mais n'ont pour objet que **l'hyper-présent ou le futur à venir**. Elles servent à comprendre le monde de demain, à le faire advenir dans de meilleures conditions et à l'expliquer aux citoyens, aux dirigeants, etc. Le chercheur en Humanités est ici dans une toute autre situation par rapport à ce qu'il a l'habitude de faire dans sa pratique. Il doit **inventer, comprendre et analyser** en s'appuyant sur une solide culture classique scientifique qui le met en position de mieux **anticiper les changements dans la société**. Tel est, à mon sens, le rôle des HN. Bien loin des outils numériques utilisés pour notre recherche dans des secteurs disciplinaires donnés... »

Source : « Les Humanités Numériques : une nouvelle discipline universitaire ? », Suzanne Dumouchel, in *Digital Humanities à l'Institut historique allemand*, publié le 19/06/2015, consulté le 05/04/2019

Humanités numériques

Concrètement :

- Édition numérique,
- Bibliothèques et expositions numériques,
- Traitement automatique des textes et images,
- Bases de données appliquées aux objets patrimoniaux,
- Archivage numérique,
- Programmation web pour institutions culturelles…

Numérisation

- Production de la représentation numérique d'un objet physique.
- « Conversion » de données depuis un support physique vers une forme et un support numériques.
- Adaptation, voire remaniement des données brutes.

Édition

1) « Action de faire paraître un texte et d'en assurer la diffusion auprès du public, directement ou par le truchement d'intermédiaires. »

2) « Ensemble des opérations intellectuelles et matérielles par lesquelles le texte d'une œuvre est établi. »

F. Duval
Les mots de l'édition de texte
Les manuels de l'École des chartes, 2015

Édition numérique

- « Édition dont les données sont encodées et accessibles numériquement. »

F. Duval
Les mots de l'édition de texte
Les manuels de l'École des chartes, 2015

- Un texte sur un support numérique et dans un format numérique, natif ou non.

Édition numérique

Pour nous :

- Travail scientifique, utilisant les moyens numériques, d'adaptation pour le public d'une œuvre rendue obscure à l'œil moderne par l'état de la langue ou de l'écriture dans lesquelles elle a été initialement conçue et/ou représentée.

- Travail d'enrichissement scientifique des données brutes d'une œuvre par des moyens numériques.

L'utilisation du numérique change la forme mais non le fond de la démarche. Elle change également les possibilités de l'éditeur.

Édition numérique

Pourquoi une édition numérique ?

- Faciliter le partage des travaux scientifiques.
- Faciliter la recherche et l'analyse automatiques.
- Profiter des nombreuses possibilités structurelles et sémantiques des langages numériques.
- Repenser l'approche d'une œuvre en s'affranchissant des contraintes du format papier.

Édition numérique

Pourquoi une édition numérique ?

- Faciliter le partage des travaux scientifiques.
- Faciliter la recherche et l'analyse automatiques.
- Profiter des nombreuses possibilités structurelles et sémantiques des langages numériques.
- Repenser l'approche d'une œuvre en s'affranchissant des contraintes du format papier.

Chaîne de numérisation

L'ensemble des procédés, physiques et numériques, permettant de passer d'un support papier originel à sa représentation numérique.

Les étapes de la numérisation

1) Travail initial.
2) Numérisation.
3) Acquisition du texte.
4) Conception de la base de données.
5) Encodage du texte.
6) Conception des outils de consultation.
7) Mise à disposition du public.
8) Maintenance.

Les étapes de la numérisation

1) Travail initial.
2) Numérisation.
3) Acquisition du texte.
4) Conception de la base de données.
5) Encodage du texte.
6) Conception des outils de consultation.
7) Mise à disposition du public.
8) Maintenance.

1) Travail initial

- Travail codicologique et bibliographique.
- Définition des buts et cibles.
- Évaluation du travail à venir.

Les étapes de la numérisation

1) Travail initial.
2) Numérisation.
3) Acquisition du texte.
4) Conception de la base de données.
5) Encodage du texte.
6) Conception des outils de consultation.
7) Mise à disposition du public.
8) Maintenance.

2) Numérisation

Acquisition d'images de la source.

En fonction de :

- Accès à la source.
- État de la source.
- Taille de la source.
- Matériel.
- Buts scientifiques et techniques.

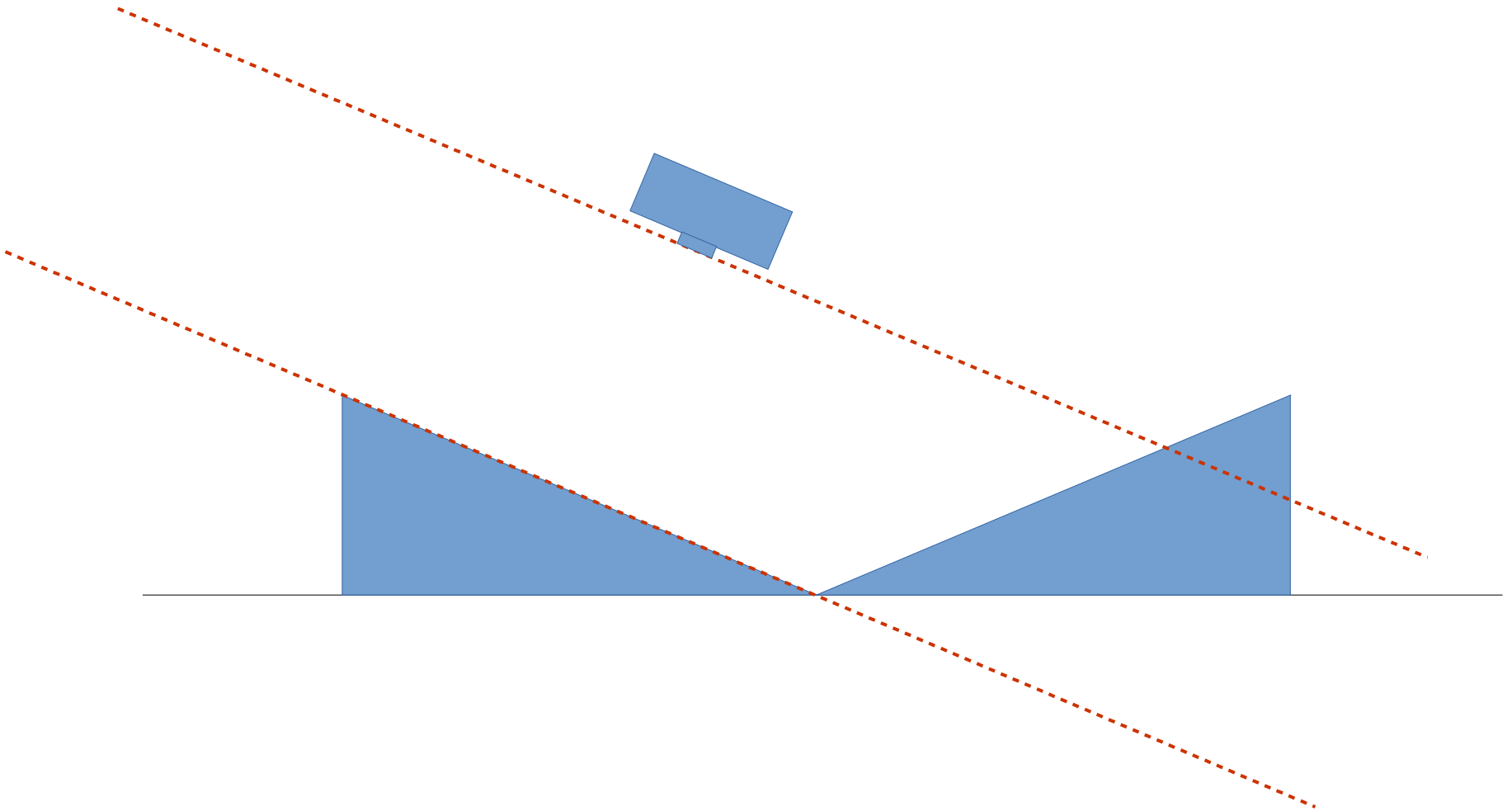
2) Numérisation

Source : p. 316, Très Ancien Coutumier, Paris, Bibliothèque Sainte-Geneviève, in ms. fr. 1743

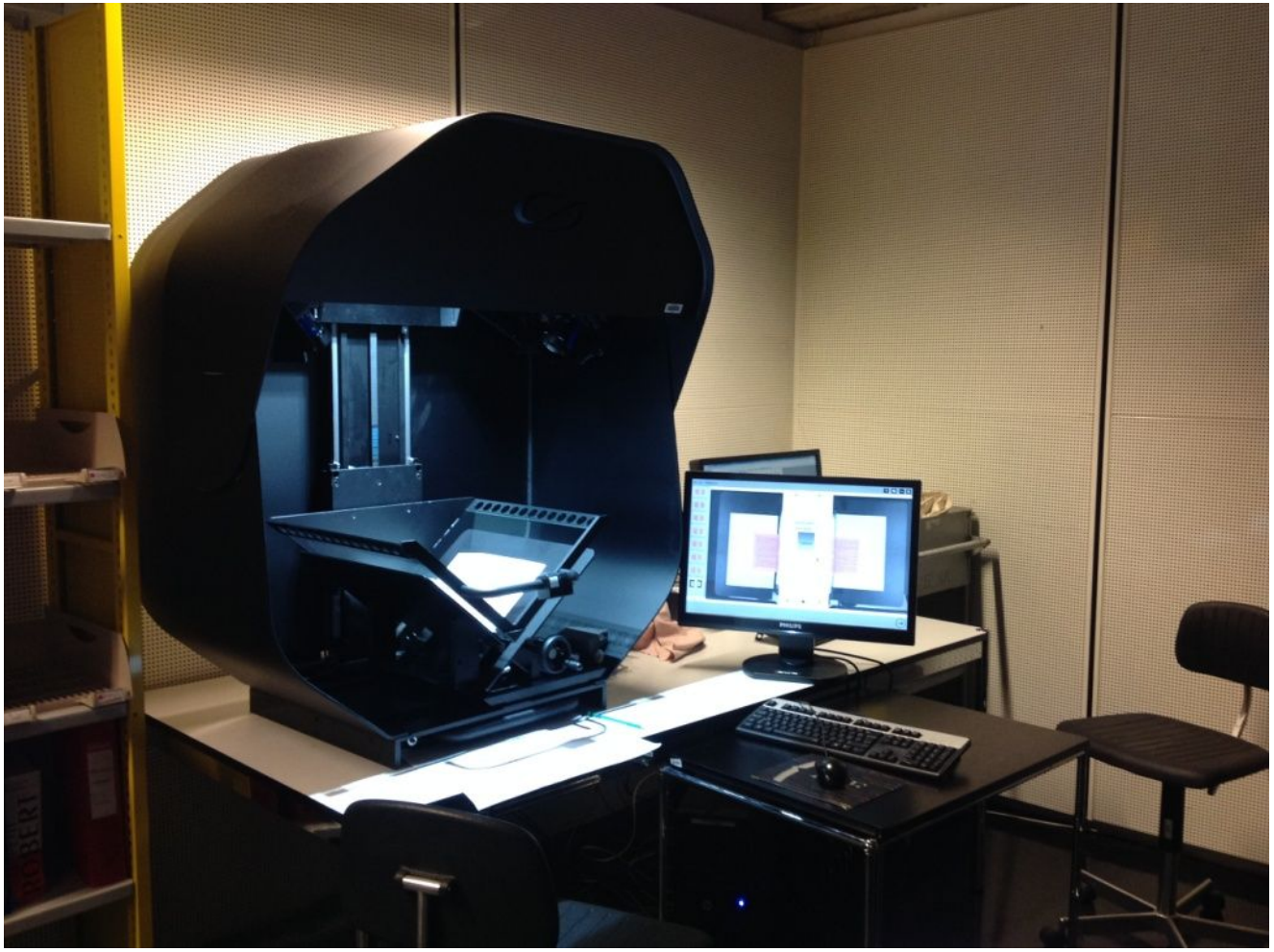
2) Numérisation

Source : Au delà des murs Créations - https://www.audeladesmurscreations.com/product-page/ref-113h-housse-pour-113-113p

2) Numérisation

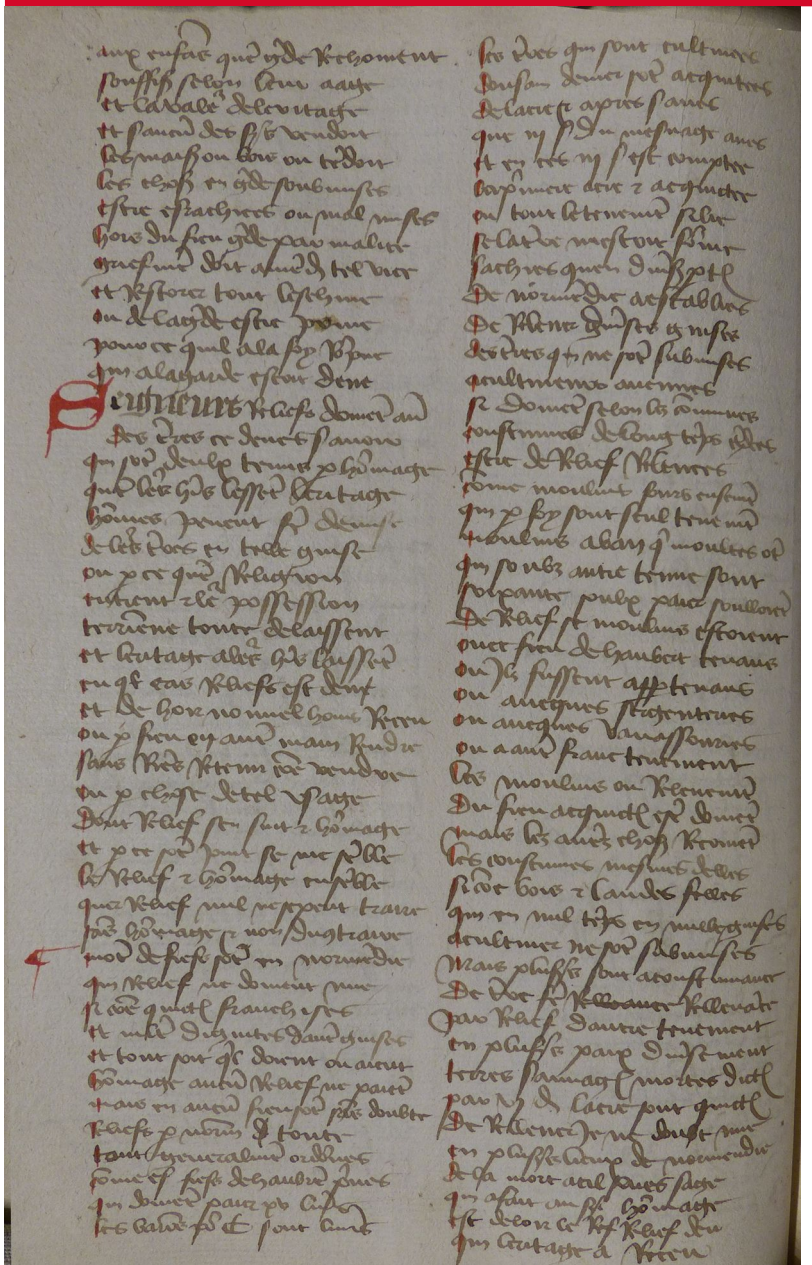


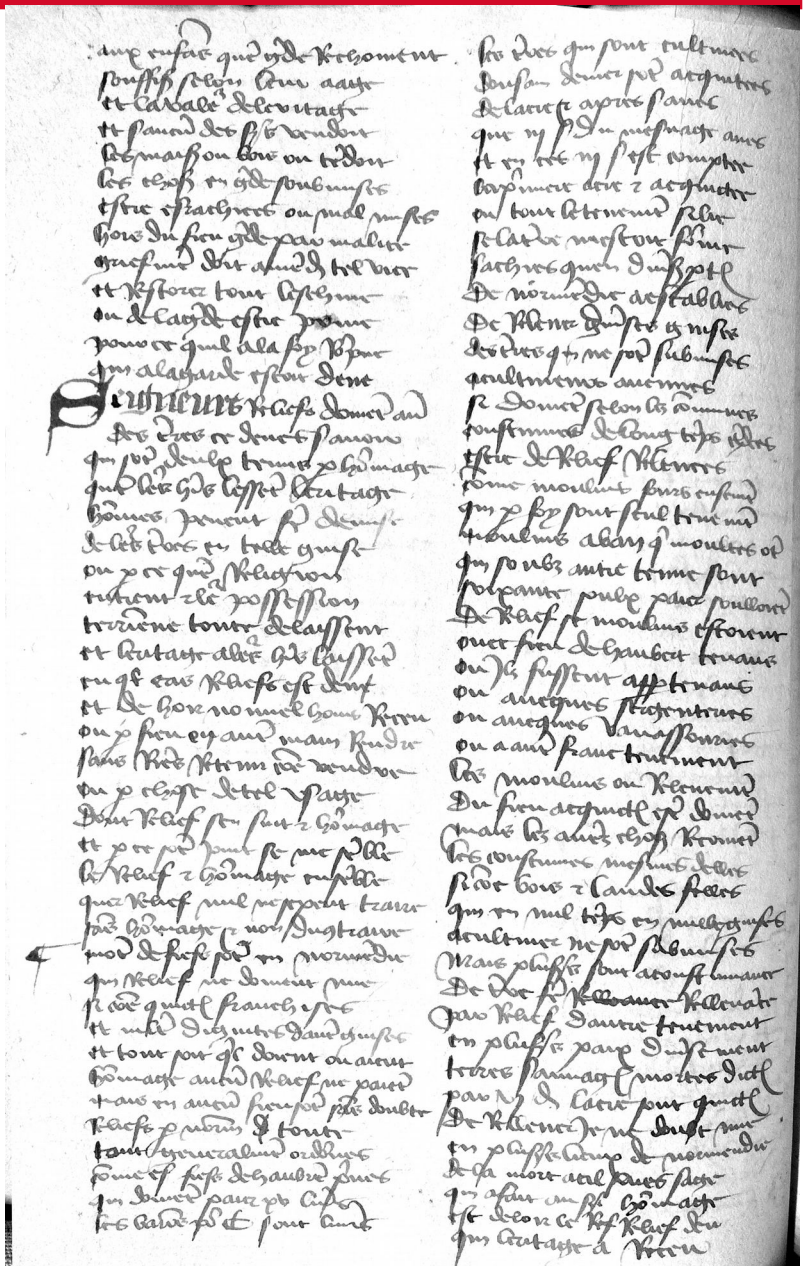
2) Numérisation



Source : « Dans les coulisses de Gallica », http://peccadille.net/2015/03/03/dans-les-coulisses-de-gallica/

2) Numérisation





Les étapes de la numérisation

1) Travail initial.
2) Numérisation.
3) Acquisition du texte.
4) Conception de la base de données.
5) Encodage du texte.
6) Conception des outils de consultation.
7) Mise à disposition du public.
8) Maintenance.

3) Acquisition du texte

- Acquisition manuelle (lecture et saisie humaines).
- Automatisation :
  - ➔ OCR (Optical Character Recognition) : caractères majoritairement prédéfinis, conçu pour les textes contemporains imprimés.
  - ➔ HTR (Handwritten Text Recognition) : apprentissage progressif du programme permettant de prendre en compte les variantes d'une même lettre dans une écriture manuscrite.
    - ➢ Sortie en texte brut.

3) Acquisition du texte

Junicode

Projet de créer des caractères Unicode
pour retranscrire les sources antiques et médiévales
dans une police gratuite et standard.

ℂ ꝛ ꝓ 2 9 ꝯ ſ ﬅ ā t̓ ꝺ ꝟ

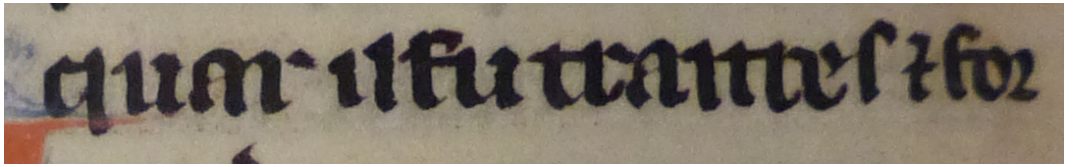
Développée par Peter Baker à l'université de Virginie
dans le cadre du projet MUFI.

Sources : « Medieval Unicode Font Initiative » - https://folk.uib.no/hnooh/mufi/
« Junicode » - http://junicode.sourceforge.net/

3) Acquisition du texte

Entraînement d'un modèle d'HTR

- Transcription initiale manuelle d'un extrait représentatif.
- Synchronisation texte-image pour le modèle (caractère par caractère).



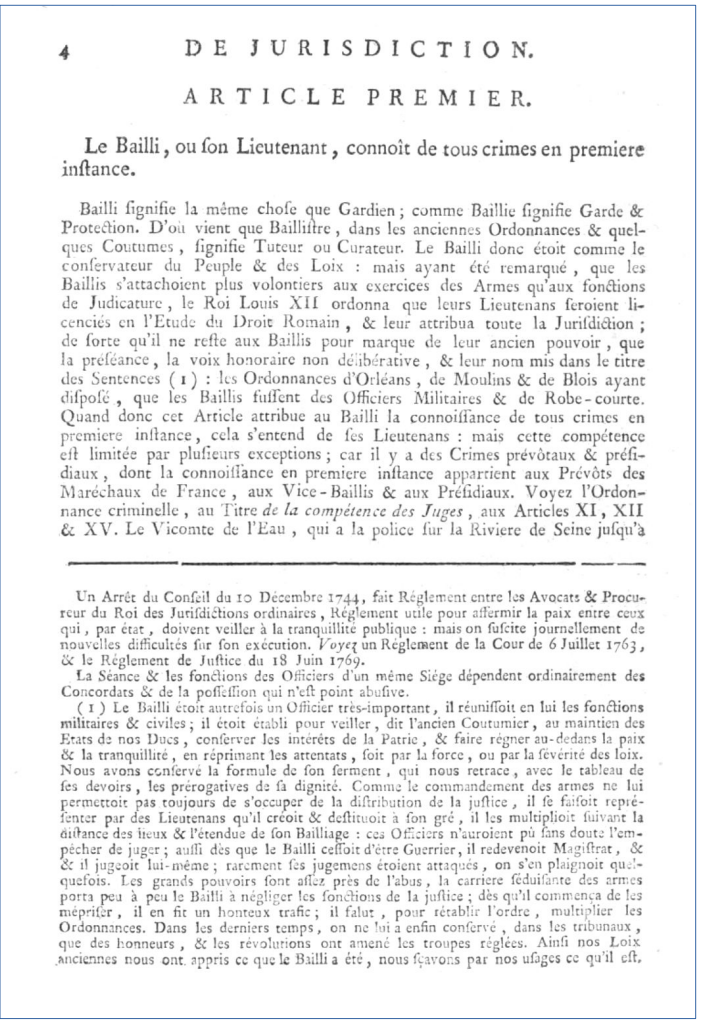
quar il fu traicteſ ⁊ fo2

3) Acquisition du texte

Entraînement d'un modèle d'HTR

- Transcription initiale manuelle d'un extrait représentatif.
- Synchronisation texte-image pour le modèle.
- Apprentissage initial suivant la synchronisation.
- Test initial sur de nouvelles pages.
- Correction du test.
- Reprise de l'apprentissage, et encore, et encore… jusqu'à satisfaction.

3) Acquisition du texte



4 DE JURISDICTION.

ARTICLE PREMIER.

Le Bailli, ou ſon Lieutenant, connoît de tous crimes en premiere inſtance.

Bailli ſignifie la même choſe que Gardien ; comme Baillie ſignifie Garde & Protection. D'où vient que Bailliſtre, dans les anciennes Ordonnances & quelques Coutumes, ſignifie Tuteur ou Curateur. Le Bailli donc étoit comme le conſervateur du Peuple & des Loix : mais ayant été remarqué, que les Baillis s'attachoient plus volontiers aux exercices des Armes qu'aux fonctions de Judicature, le Roi Louis XII ordonna que leurs Lieutenans ſeroient licenciés en l'Etude du Droit Romain, & leur attribua toute la Juriſdiction ; de ſorte qu'il ne reſte aux Baillis pour marque de leur ancien pouvoir, que la préſéance, la voix honoraire non délibérative, & leur nom mis dans le titre des Sentences (1) : les Ordonnances d'Orléans, de Moulins & de Blois ayant diſpoſé, que les Baillis fuſſent des Officiers Militaires & de Robe-courte. Quand donc cet Article attribue au Bailli la connoiſſance de tous crimes en premiere inſtance, cela s'entend de ſes Lieutenans : mais cette compétence eſt limitée par pluſieurs exceptions ; car il y a des Crimes prévôtaux & préſidiaux, dont la connoiſſance en premiere inſtance appartient aux Prévôts des Maréchaux de France, aux Vice-Baillis & aux Préſidiaux. Voyez l'Ordonnance criminelle, au Titre *de la compétence des Juges*, aux Articles XI, XII & XV. Le Vicomte de l'Eau, qui a la police ſur la Riviere de Seine juſqu'à

Un Arrêt du Conſeil du 10 Décembre 1744, fait Réglement entre les Avocats & Procureur du Roi des Juriſdictions ordinaires, Réglement utile pour affermir la paix entre ceux qui, par état, doivent veiller à la tranquillité publique : mais on ſuſcite journellement de nouvelles difficultés ſur ſon exécution. *Voyez* un Réglement de la Cour de 6 Juillet 1763, & le Réglement de Juſtice du 18 Juin 1769.

La Séance & les fonctions des Officiers d'un même Siége dépendent ordinairement des Concordats & de la poſſeſſion qui n'eſt point abuſive.

(1) Le Bailli étoit autrefois un Officier très-important, il réuniſſoit en lui les fonctions militaires & civiles ; il étoit établi pour veiller, dit l'ancien Coutumier, au maintien des Etats de nos Ducs, conſerver les intérêts de la Patrie, & faire régner au-dedans la paix & la tranquillité, en réprimant les attentats, ſoit par la force, ou par la ſévérité des loix. Nous avons conſervé la formule de ſon ſerment, qui nous retrace, avec le tableau de ſes devoirs, les prérogatives de ſa dignité. Comme le commandement des armes ne lui permettoit pas toujours de s'occuper de la diſtribution de la juſtice, il ſe faiſoit repréſenter par des Lieutenans qu'il créoit & deſtituoit à ſon gré, il les multiplioit ſuivant la diſtance des lieux & l'étendue de ſon Bailliage : ces Officiers n'auroient pû ſans doute l'empêcher de juger ; auſſi dès que le Bailli ceſſoit d'être Guerrier, il redevenoit Magiſtrat, & il jugeoit lui-même ; rarement ſes jugemens étoient attaqués, on s'en plaignoit quelquefois. Les grands pouvoirs ſont aſſez près de l'abus, la carriere ſéduiſante des armes porta peu à peu le Bailli à négliger les fonctions de la juſtice ; dès qu'il commença de les mépriſer, il en fit un honteux trafic ; il falut, pour rétablir l'ordre, multiplier les Ordonnances. Dans les derniers temps, on ne lui a enfin conſervé, dans les tribunaux, que des honneurs, & les révolutions ont amené les troupes réglées. Ainſi nos Loix anciennes nous ont appris ce que le Bailli a été, nous ſçavons par nos uſages ce qu'il eſt.

4
DE JURISDICTION.
ARTICLE PREMIER.
Le Bailli, ou ſon Lieutenant, connoît de tous crimes en premiere inſtance.
Bailli ſignifie la même choſe que Gardien ; comme Baillie ſignifie Garde & Protection. D'où vient que Bailliſtre , dans les anciennes Ordonnances & quelques Coutumes , ſignifie Tuteur ou Curateur. Le Bailli donc étoit comme le conſervateur du Peuple & des Loix : mais ayant été remarqué , que les Baillis s'attachoient plus volontiers aux exercices des Armes qu'aux fonctions de Judicature , le Roi Louis XII ordonna que leurs Lieutenans ſeroient licenciés en l'Etude du Droit Romain , & leur attribua toute la Juriſdiction ; de ſorte qu'il ne reſte aux Baillis pour marque de leur ancien pouvoir , que la préſéance , la voix honoraire non délibérative , & leur nom mis dans le titre des Sentences (I) : les Ordonnances d'Orléans , de Moulins & de Blois ayant diſposé , que les Baillis fuſſent des Officiers Militaires & de Robe-courte. Quand donc cet Article attribue au Bailli la connoiſſance de tous crimes en premiere inſtance, cela s'entend de ſes Lieutenans : mais cette compétence eſt limitée par pluſieurs exceptions ; car il y a des Crimes prévôtaux & préſidiaux, dont la connoiſſance en premiere inſtance appartient aux Prévôts des Maréchaux de France, aux Vice-Baillis & aux Préſidiaux. Voyez l'Ordonnance criminelle, au Titre de la compétence des Juges, aux Articles XI, XII & XV. Le Vicomte de l'Eau , qui a la police ſur la Riviere de Seine juſqu'à
Un Arrêt du Conſeil du 10 Décembre 1744, fait Réglement entre les Avocats & ous ont appris ce que le Bailli a été, nous ſçavons par nos uſages ce qu'il eſt.

Les étapes de la numérisation

1) Travail initial.
2) Numérisation.
3) Acquisition du texte.
4) Conception de la base de données.
5) Encodage du texte.
6) Conception des outils de consultation.
7) Mise à disposition du public.
8) Maintenance.

4) Conception de la base de données

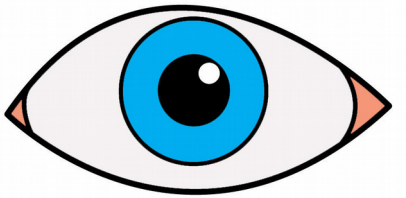
- Base de données textuelle, très souvent en XML-TEI (grandes et vastes possibilités et utilisation très large garante d'interopérabilité et pérennité).
- À nouveau, quelle finalité ?
- Quelle serait la structure logique du *corpus* ?
- Quel temps de travail a-t-on ?
- Quels outils sont prévus ?

Les étapes de la numérisation

1) Travail initial.
2) Numérisation.
3) Acquisition du texte.
4) Conception de la base de données.
5) Encodage du texte.
6) Conception des outils de consultation.
7) Mise à disposition du public.
8) Maintenance.

5) Encodage du texte - XML

Texte



- Police plus grande
- Placé au-dessus
- Séparé du corps

= titre

Un **texte** est une série orale ou écrite de mots perçus comme constituant un ensemble cohérent, porteur de sens et utilisant les structures propres à une langue (conjugaisons, construction et association des phrases...)1. Un texte n'a pas de longueur déterminée sauf dans le cas de poèmes à forme fixe comme le sonnet ou le haïku.

L'étude formelle des textes s'appuie sur la linguistique, qui est l'approche scientifique du langage.

5) Encodage du texte - XML

Texte

Chaîne de caractères,

- Police : « Liberation Sans »
- Taille : 22 px.
- Couleur : noire.
- Alignement : gauche.

5) Encodage du texte - XML

Chaîne de caractères

- “Texte”

Enrichissement structurel

- = titre

Enrichissement formel

- police
- couleur
- position…

5) Encodage du texte - XML

Enrichissement structurel

- = titre

XML

5) Encodage du texte - XML

- = *eXtensible Markup Language*
- Langage d'encodage (≠ programmation)
- Grammaire : langage à balises.

<titre>Texte</titre>

<titre type="principal">Texte</titre>

- Finesse de l'encodage jusqu'au caractère.
- « eXtensible » = mélange de vocabulaires XML

5) Encodage du texte - TEI

- = *Text Encoding Initiative*

https://tei-c.org

- Nom de domaine = vocabulaire spécifique
  = set de balises prédéfinies

- Encodage sémantique et structurel des textes oraux et écrits, quel que soit le support originel.
- Description fine et précise des supports originels et métadonnées.

5) Encodage du texte - TEI

toy cat has bell in tail which continues to make a tinkling sound
Mar meows

Ros – because it is so old
Mar – how ‘bout *your* cat… yours is *new*

shows Father the cat

```
<incident>
    <desc>toy cat has bell in tail which continues to make a
        tinkling sound</desc>
</incident>
<vocal who="#mar">
    <desc>meows</desc>
</vocal>
<u who="#ros">because it is so old</u>
<u who="#mar">how
    <choice>
        <orig>bout</orig>
        <reg>about</reg>
    </choice>
    <emph>your</emph> cat
    <pause/>yours is
    <emph>new</emph>
    <kinesic>
        <desc>shows Father the cat</desc>
    </kinesic>
</u>
```

5) Encodage du texte - TEI

Document daté un 10 avril (par le scribe)
entre 851 et 867 (par l'éditeur).

```
<date>
    <date when="--04-10">
        un 10 avril,
    </date>
    <date notBefore="0851" notAfter="0867" cert="high">
        entre 851 et 867
    </date>
</date>
```

5) Encodage du texte - TEI

Sujet

Coutume

Commentaire français

Commentaire latin

Entête

Texte principal (gauche, droite)

Glose encadrante (gauche, droite)

5) Encodage du texte - TEI

```
<incident>
    <desc>toy cat has bell in tail which continues to make a
        tinkling sound</desc>
</incident>
<vocal who="#mar">
    <desc>meows</desc>
</vocal>
<u who="#ros">because it is so old</u>
<u who="#mar">how
    <choice>
        <orig>bout</orig>
        <reg>about</reg>
    </choice>
    <emph>your</emph> cat
    <pause/>yours is
    <emph>new</emph>
    <kinesic>
        <desc>shows Father the cat</desc>
    </kinesic>
</u>
```

toy cat has bell in tail which continues to make a tinkling sound
Mar meows

Ros – because it is so old
Mar – how ‘bout *your* cat… yours is *new*

shows Father the cat

Les étapes de la numérisation

1) Travail initial.
2) Numérisation.
3) Acquisition du texte.
4) Conception de la base de données.
5) Encodage du texte.
6) Conception des outils de consultation.
7) Mise à disposition du public.
8) Maintenance.

7) Outils de consultation

Site internet :

- Serveur,
- Base de données (= fichiers XML-TEI)
- Système d'interrogation (pour naviguer dans le XML-TEI)
- Transformations (pour l'affichage du XML),
- Structure HTML/etc.

Les étapes de la numérisation

1) Travail initial.
2) Numérisation.
3) Acquisition du texte.
4) Conception de la base de données.
5) Encodage du texte.
6) Conception des outils de consultation.
7) Mise à disposition du public.
8) Maintenance.

Les étapes de la numérisation

1) Travail initial.
2) Numérisation.
3) Acquisition du texte.
4) Conception de la base de données.
5) Encodage du texte.
6) Conception des outils de consultation.
7) Mise à disposition du public.
8) Maintenance.

8) Maintenance

/!\ TRÈS IMPORTANT /!\

- Pérennité des outils et de la base de données.