Recueil et structuration de corpus – TD 4

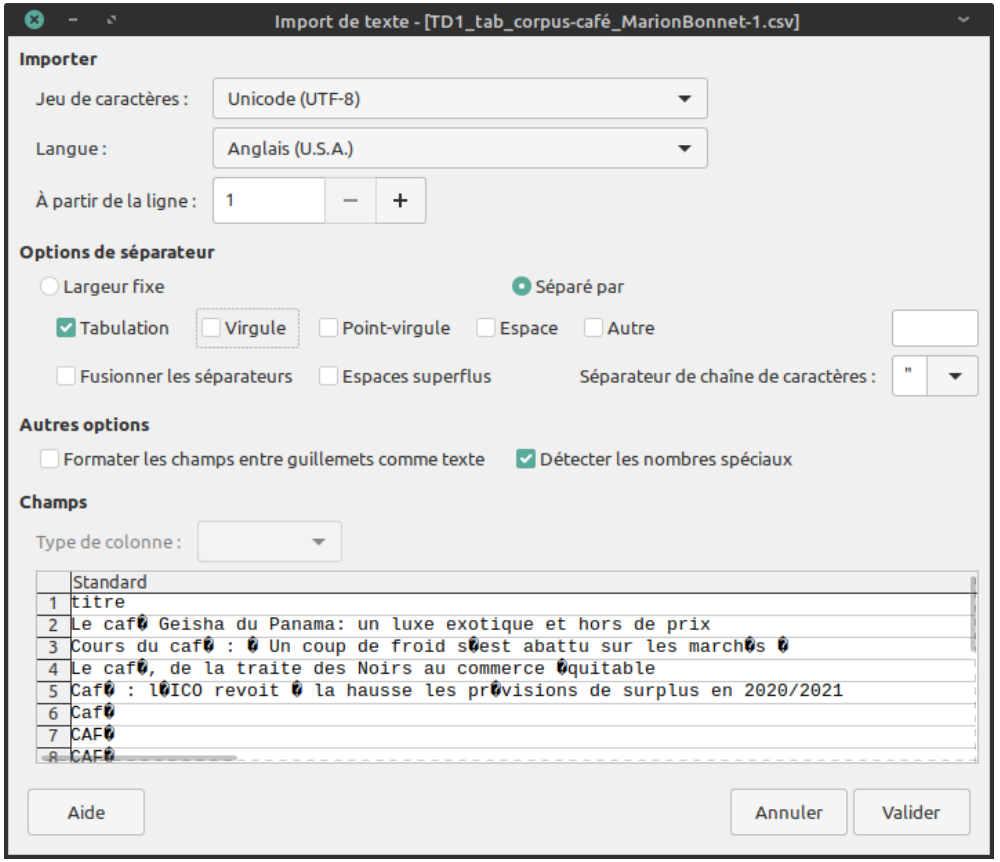
Achille Falaise – Alexandre Roulois

Plan du TD

- Correction TD
- Édition XML avancée
- Langage XPath
- Collecte de documents Web (*Web Scraping*)
  - Approche manuelle
  - Approche automatique
- Importation d’un corpus dans TXM

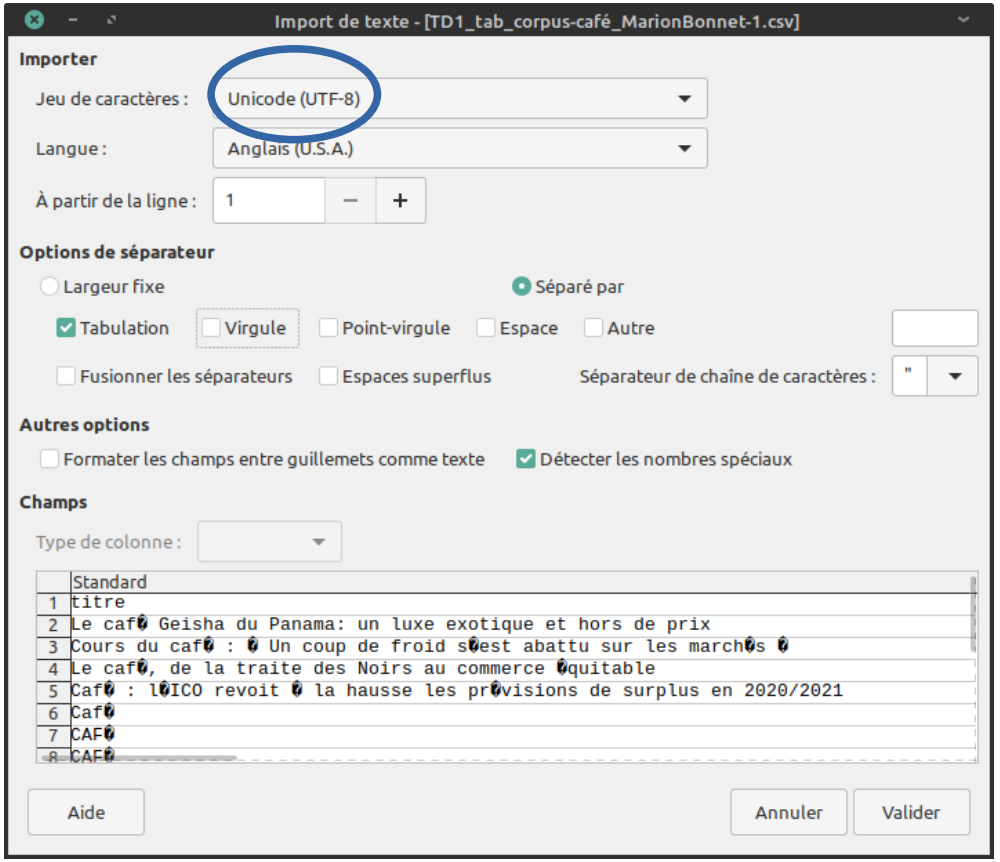
Correction

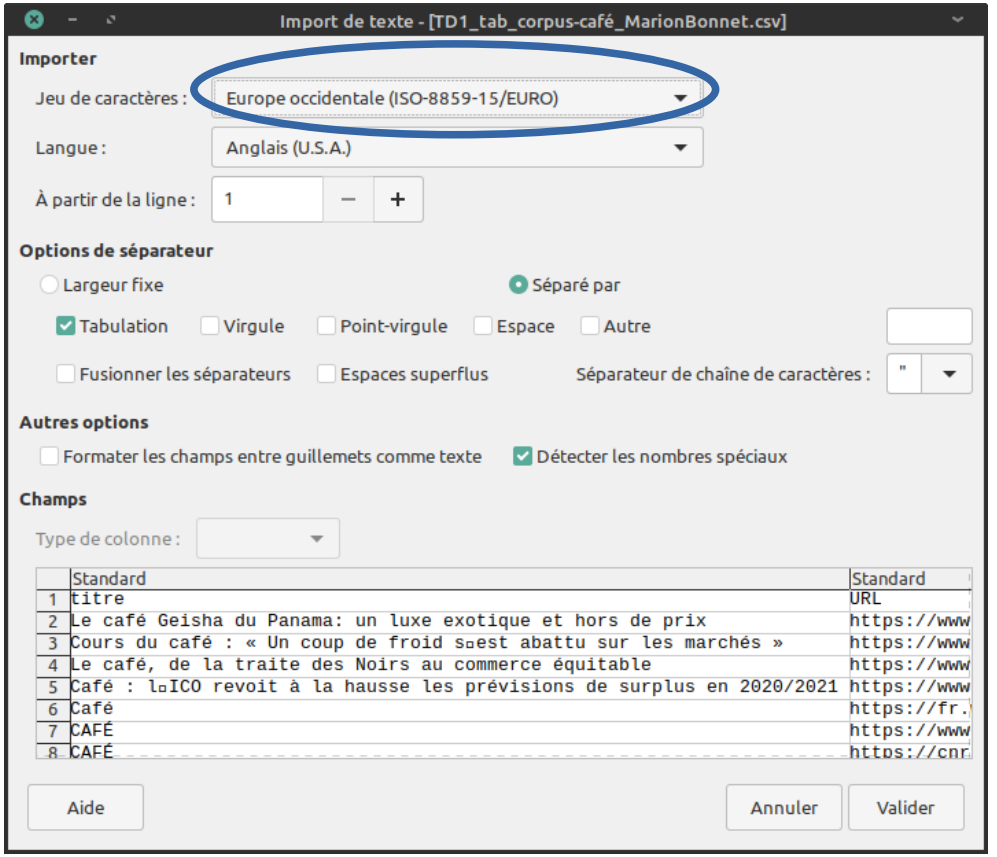
- Cherchez l’erreur !



Correction

- Cherchez l’erreur !





Correction

- Cherchez le problème !

	https://www.lemonde.fr/economie/article/2013/09/27/le-cafe-[illegible]	article de presse	U
en	https://www.agenceecofin.com/cafe/0302-84798-cafe-l-ico-re	article de presse	U
	https://fr.wikipedia.org/wiki/Caf%C3%A9	texte encyclopédique	U
	https://www.universalis.fr/encyclopedie/cafe/	texte encyclopédique	U
	https://cnrtl.fr/definition/caf%C3%A9	texte encyclopédique	U
	https://www.larousse.fr/dictionnaires/francais/caf%C3%A9/1	texte encyclopédique	U
	https://www.researchgate.net/profile/Hayat_Zirari/publicatio	texte scientifique	U
man	https://www.sciencedirect.com/science/article/abs/pii/S000	texte scientifique	U
aluat	file:///C:/Users/UTILIS~1/AppData/Local/Temp/2530-8388-1-	texte scientifique	U
	http://www.minerva-ebm.be/fr/article/756	texte scientifique	U
	https://leguidedubarista.com/guide-acidite-cafe/	blog	U
tible	https://www.maxicoffee.com/blog/lavazza-capsules-compos	blog	U
péc	https://www.latelierdescafes.com/abonnement-cafe-grain-mo	blog	U
blo	https://www.maathiildee.com/le-coffee-shop-ma-nouvelle-hal	blog	U

Correction

- Cherchez le problème !

	https://www.lemonde.fr/economie/article/2013/09/27/le-cafe-[illegible]	article de presse	U
en	https://www.agenceecofin.com/cafe/0302-84798-cafe-l-ico-re	article de presse	U
	https://fr.wikipedia.org/wiki/Caf%C3%A9	texte encyclopédique	U
	https://www.universalis.fr/encyclopedie/cafe/	texte encyclopédique	U
	https://cnrtl.fr/definition/caf%C3%A9	texte encyclopédique	U
	https://www.larousse.fr/dictionnaires/francais/caf%C3%A9/1	texte encyclopédique	U
	https://www.researchgate.net/profile/Hayat_Zirari/publicatio	texte scientifique	U
man	https://www.sciencedirect.com/science/article/abs/pii/S000	texte scientifique	U
alua	file:///C:/Users/UTILIS~1/AppData/Local/Temp/2530-8388-1-	texte scientifique	U
	http://www.minerva-ebm.be/fr/article/756	texte scientifique	U
	https://leguidedubarista.com/guide-acidite-cafe/	blog	U
tible	https://www.maxicoffee.com/blog/lavazza-capsules-compos	blog	U
péc	https://www.latelierdescafes.com/abonnement-cafe-grain-mo	blog	U
	https://www.maathiildee.com/le-coffee-shop-ma-nouvelle-hal	blog	

Correction

- Un choix de balise qui pourrait être amélioré

```
  <text>
    <body>
      <p> Depuis le café du matin jusqu'au décaféiné du soir en passant par la ou les pauses café
de la journée, la boisson de café rythme la vie de centaines de millions de personnes. Dans le même
temps, la prospérité de régions entières du monde tropical est étroitement dépendante de la
production et de la commercialisation de café dans un contexte où la volatilité du cours mondial
est devenue de plus en plus marquée comme pour la plupart des autres grands produits agricoles.</p>
      <title> Les origines du café </title>
      <p> Le caféier (genre Coffea L., famille des Rubiaceae) est un petit arbre de 5 à 7 mètres de
hauteur cultivé pour ses fruits, les « cerises », qui renferment deux graines entourées de pulpe.
Ce sont ces dernières qui, après torréfaction (calcination partielle), vont servir à produire du
café. Les deux espèces les plus répandues sont Coffea arabica L., le caféier commun, originaire
d'Éthiopie et aujourd'hui très largement cultivé en Amérique centrale, en Amérique du Sud et au
Kenya, et Coffea canephora Pierre, le caféier congo, dont la variété robusta est très présente en
Afrique et en Asie. </p>
      <p> Croissant en zone intertropicale, le caféier réclame une forte humidité (de 1 à 2 mètres
de précipitations annuelles) et des températures moyennes de 20 à 25 0C. Toutefois, si Coffea
canephora est sensible aux variations de chaleur et ne peut être cultivé qu'à des altitudes pas
trop élevées, Coffea arabica peut s'accommoder de températures plus basses et de climats
d'altitude. Mais il ne supporte pas le gel, comme l'ont souligné les fortes baisses de récolte
périodiquement enregistrées au Brésil, baisses qui à chaque fois ont engendré une envolée du cours
mondial. </p>
    </body>
  </text>
```

Correction

- Attention à la balise auto-fermante : </lb> vs <lb/>

```
15   <text>
16     <body>
17       <head> Généralités </head>
18         <p> Depuis le café du matin jusqu'au décaféiné du soir en passant par la ou les pauses café de la
     journée, la boisson de café rythme la vie de centaines de millions de personnes. Et comme l'avoue Georges
     Courteline, <q> on change plus facilement de religion que de café </q>. Dans le même temps, la prospérité
     de régions entières du monde tropical est étroitement dépendante de la production et de la
     commercialisation de café dans un contexte où la volatilité du cours mondial est devenue de plus en plus
     marquée comme pour la plupart des autres grands produits agricoles. Pour plus d'iformations sur le sujet,
     consultez <titre xml:lang="fra"> L'encyclopédie du Café </titre> </p>
19       <head> Les origines du café </head>
20         <div type="sous-titre"> Botanique </div>
21           <p> Le caféier (genre Coffea L., famille des Rubiaceae) est un petit arbre de 5 à 7 mètres de
     hauteur cultivé pour ses fruits, les « cerises », qui renferment deux graines entourées de pulpe. Ce sont
     ces dernières qui, après torréfaction (calcination partielle), vont servir à produire du café. Les deux
     espèces les plus répandues sont <foreign xml:lang="lat"> Coffea arabica L. </foreign> , le caféier commun,
     originaire d'Éthiopie et aujourd'hui très largement cultivé en Amérique centrale, en Amérique du Sud et au
     Kenya, et <foreign xml:lang="lat"> Coffea canephora </foreign> Pierre, le caféier congo, dont la variété
     robusta est très présente en Afrique et en Asie. </p>
22         <div type="sous-titre"> climat et géographie </div>
23           <p> Croissant en zone intertropicale, le caféier réclame une forte humidité (de 1 à 2 mètres de
     précipitations annuelles) et des températures moyennes de 20 à 25 0C. Toutefois, si Coffea canephora est
     sensible aux variations de chaleur et ne peut être cultivé qu'à des altitudes pas trop élevées, Coffea
     arabica peut s'accommoder de températures plus basses et de climats d'altitude. Mais il ne supporte pas le
     gel, comme l'ont souligné les fortes baisses de récolte périodiquement enregistrées au Brésil, baisses qui
     à chaque fois ont engendré une envolée du cours mondial. </p>
24       <head> Quelques grands crus </head>
25         <div type="sous-titre"> Amérique du Sud </div>
26           <eg xml:lang="spa"> Colombia Supremo </lb> Tarrazu Amapola </lb> Huehuetenango La Capellania </eg
27     </body>
28   </text>
```

Correction

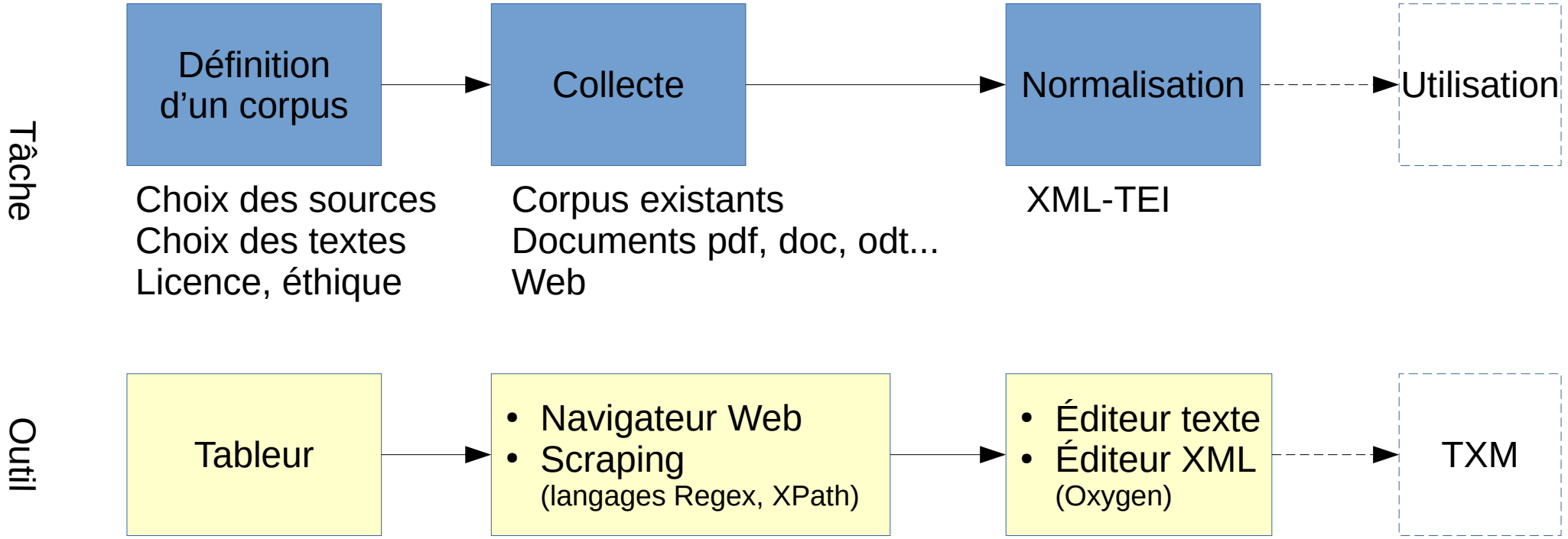
```
<p>
Le <foreign xml:lang="de">putsch</foreign> du 1er février a mis un terme à la fragile transition démocratique en
<lb/>
cours depuis dix ans. Les militaires ont instauré l'état d'urgence pour un an, arrêté
<lb/>
Aung San Suu Kyi, ainsi que des dizaines de responsables politiques et des
<lb/>
activistes. Des appels à <q> la désobéissance civile </q> avaient été lancés dès les
<lb/>
premières heures ayant suivi le coup d'Etat. En réponse, l'armée avait ordonné
<lb/>
aux fournisseurs d'accès de bloquer <term xml:lang="en">Facebook</term> et d'autres réseaux sociaux, avant
<lb/>
que les connexions ne soient partiellement rétablies dimanche.
</p>
```

```
<p>
Le vaccin d'AstraZeneca contre le <foreign xml:lang="en">Covid-19</foreign>, le troisième approuvé en France, a
<lb/>
commencé à être distribué samedi à travers le pays, afin d'être administré, en
<lb/>
premier lieu, aux professionnels des secteurs de la santé et du médico-social âgés
<lb/>
de moins de 65 ans, puis, dans un deuxième temps, aux personnes âgées de 50 à
<lb/>
64 ans. La Haute Autorité de santé (HAS) avait recommandé mardi de ne pas
<lb/>
administrer le vaccin d'AstraZeneca aux personnes de plus de 65 ans, considérant
<lb/>
qu'il <q> manque des données </q> pour cette catégorie d'âge.
</p>
```

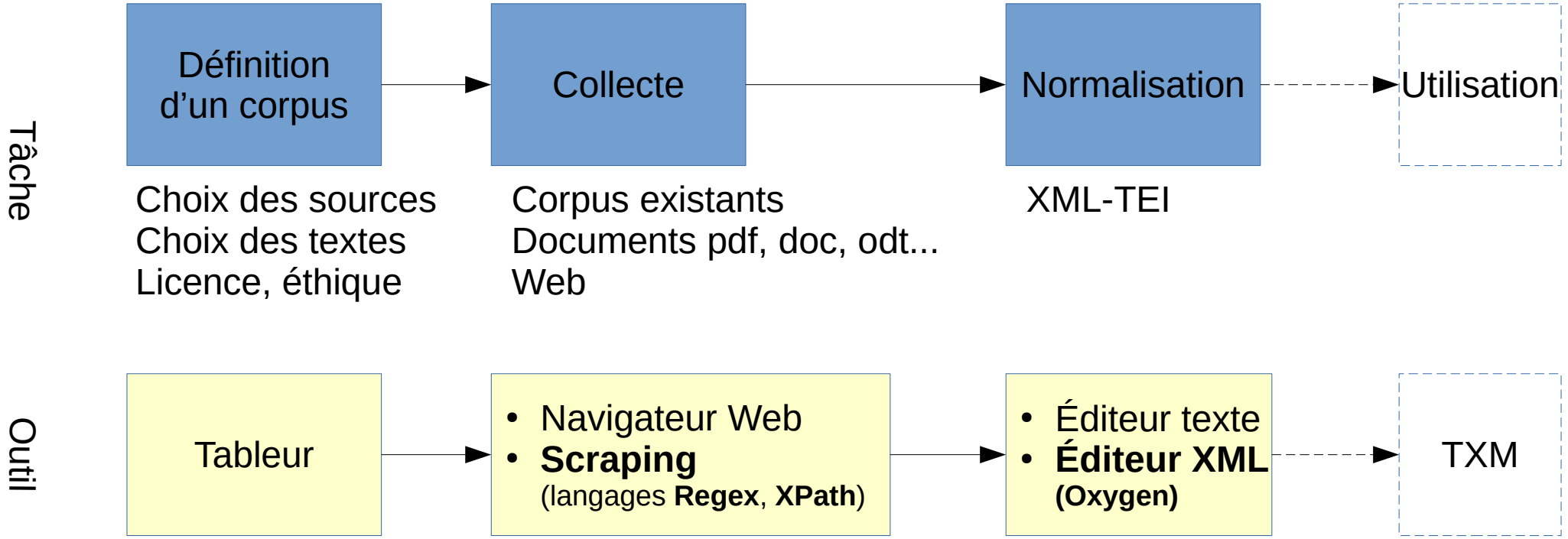
Correction

```
<eg>Chine.</eg>L’autorité chinoise de régulation des médicaments a donné son accord
<lb/>
<q> conditionnel </q> samedi pour un deuxième vaccin contre le <mentioned xml:lang="en">Covid-19</mentioned>, le
<lb/>
CoronaVac de Sinovac, fabriqué sur place. La Chine n’avait jusque-là
<lb/>
formellement approuvé qu’un seul de ses vaccins, fin décembre, celui mis au
<lb/>
point par le laboratoire Sinopharm.
</p>
```

Où en sommes nous ?

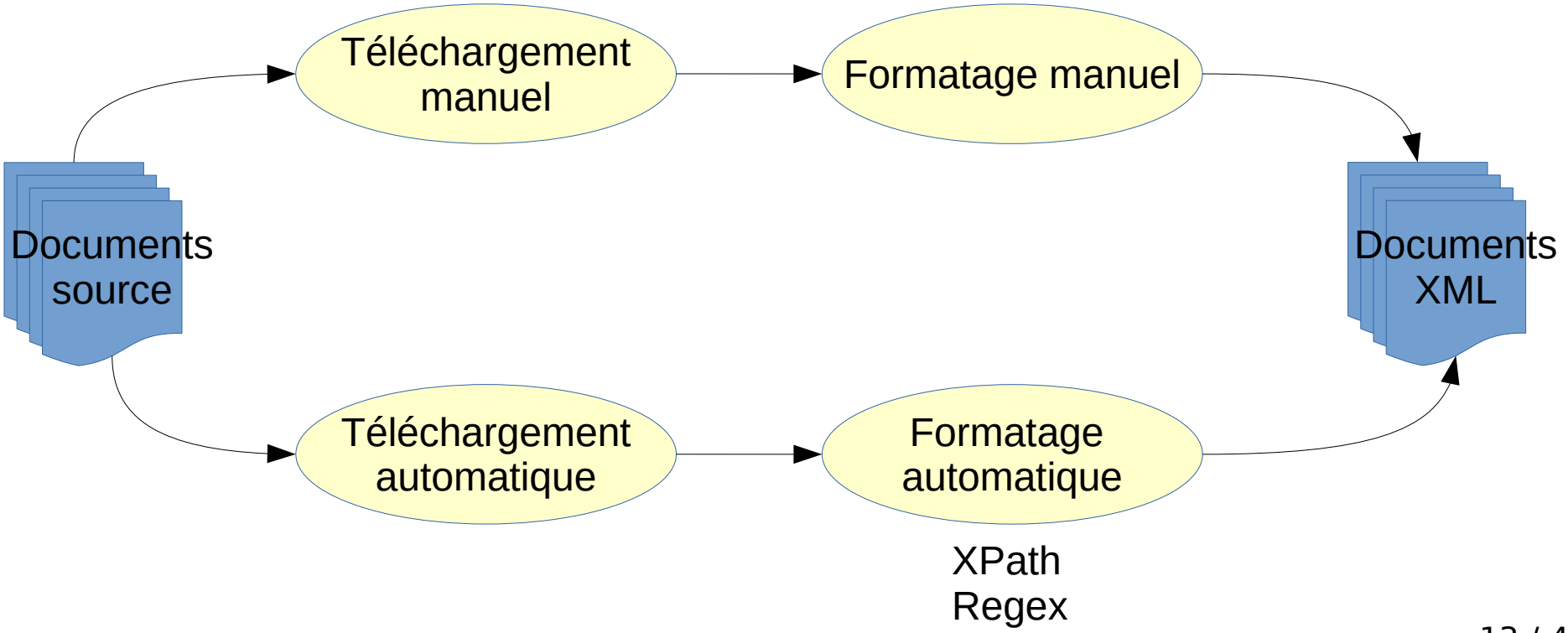


Où en sommes nous ?



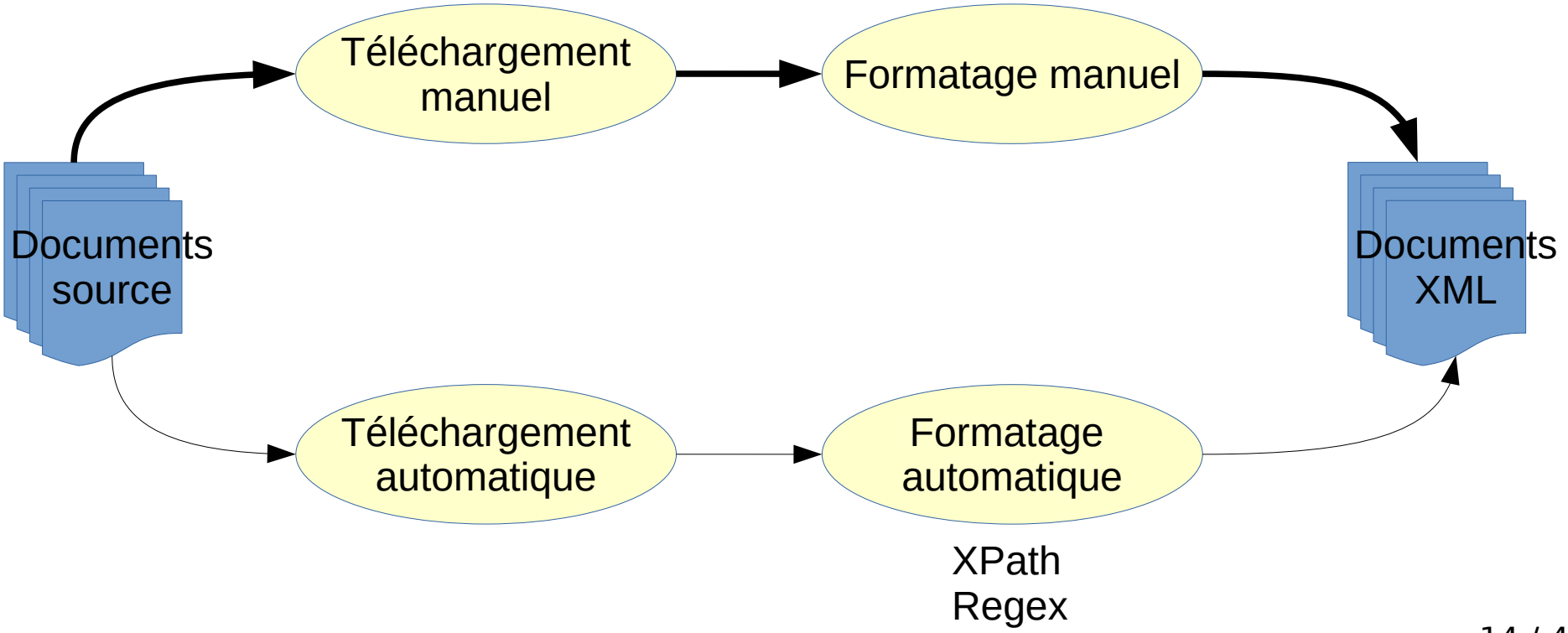
Collecte de documents Web

- *Scraping* manuel / automatisé



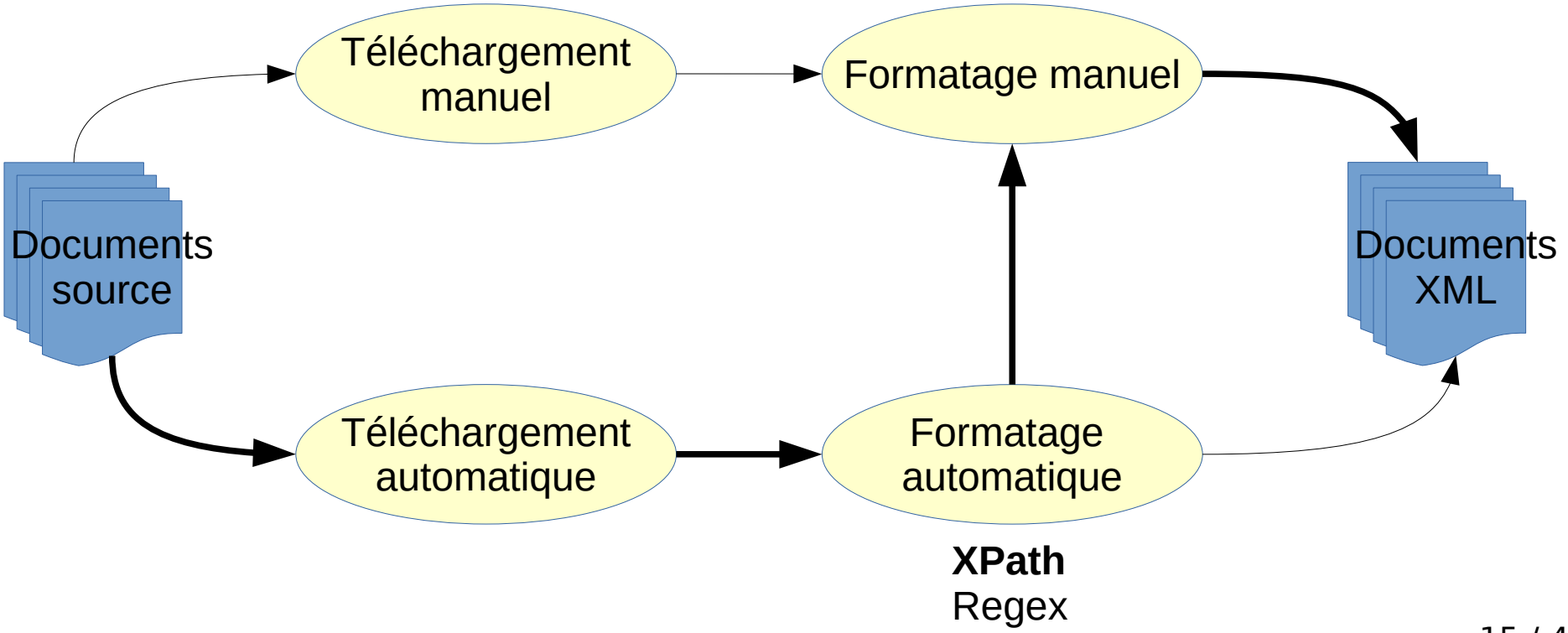
Collecte de documents Web

- *Scraping* manuel / automatisé



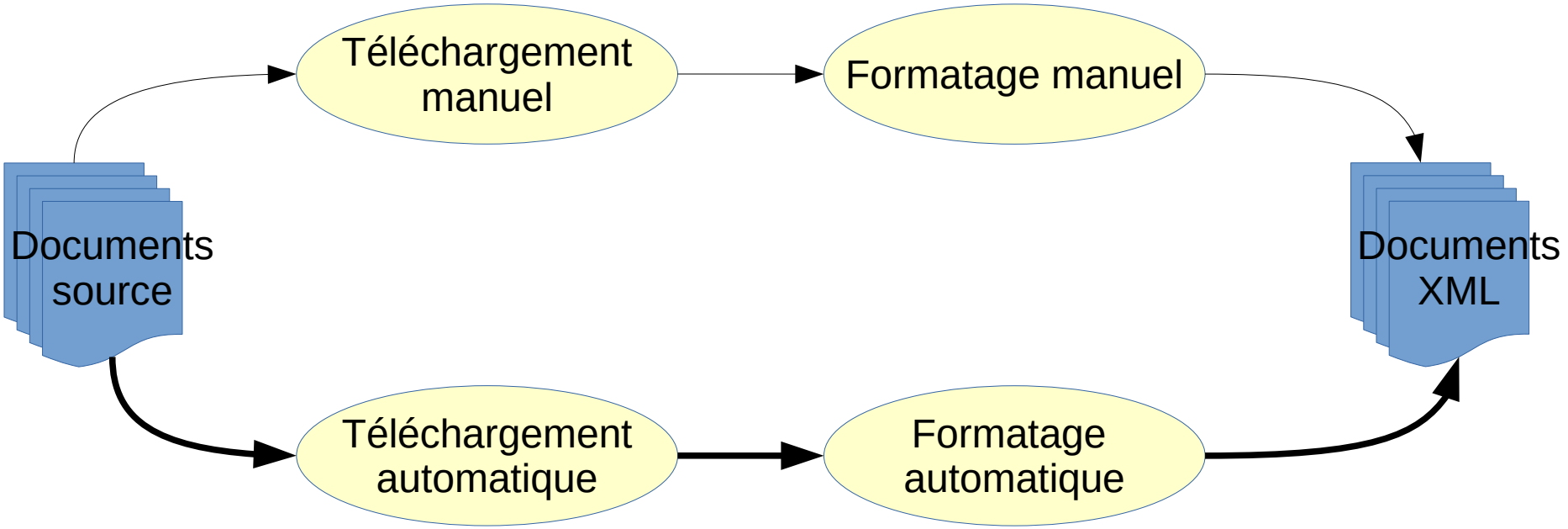
Collecte de documents Web

- *Scraping* manuel / automatisé



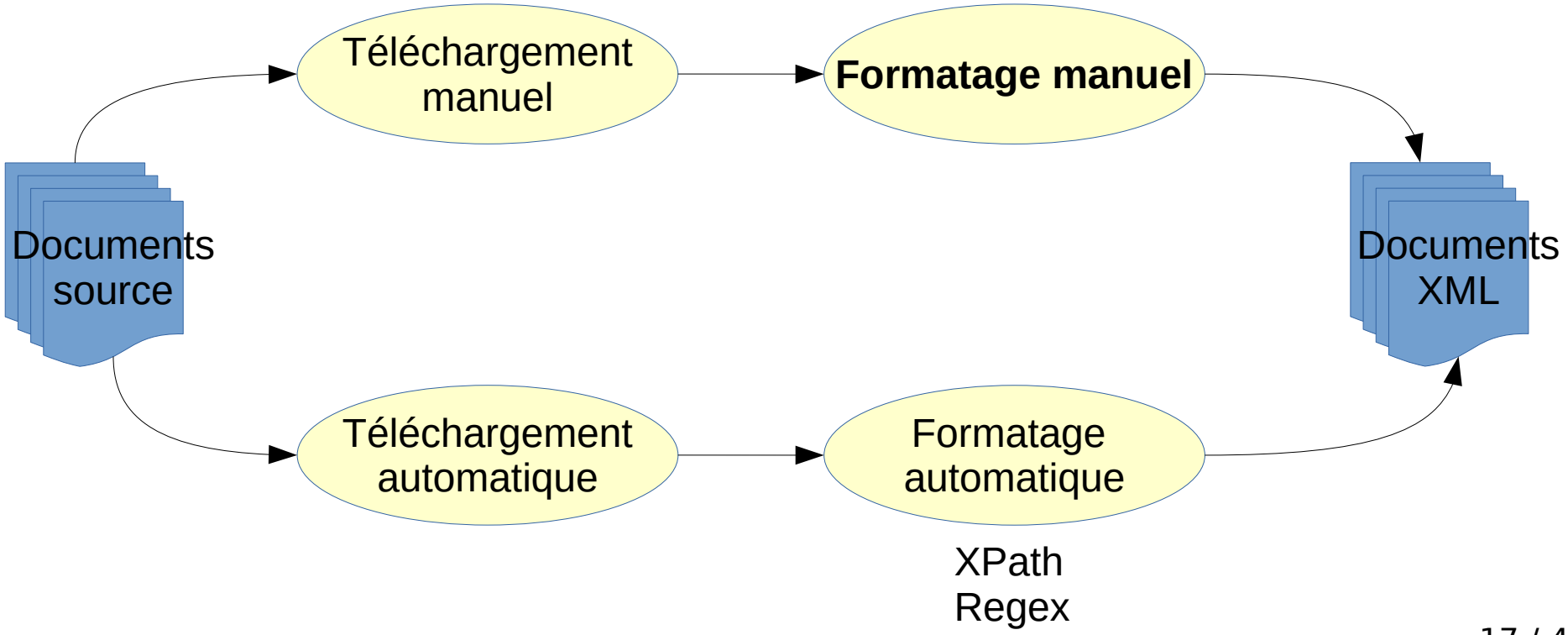
Collecte de documents Web

- *Scraping* manuel / automatisé



Édition XML

- *Scraping* manuel / automatisé

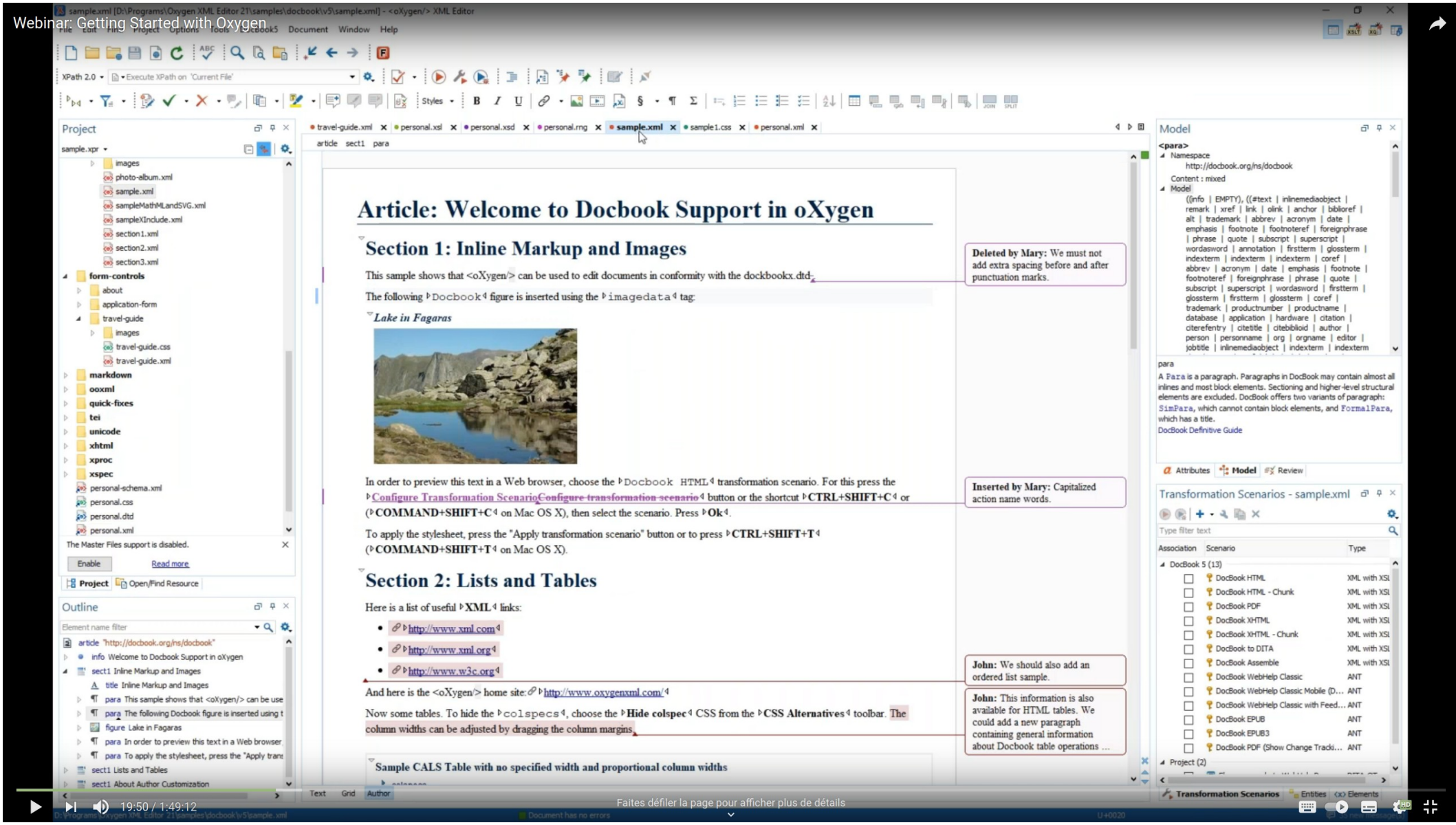


Oxygen XML Editor

New Licenses

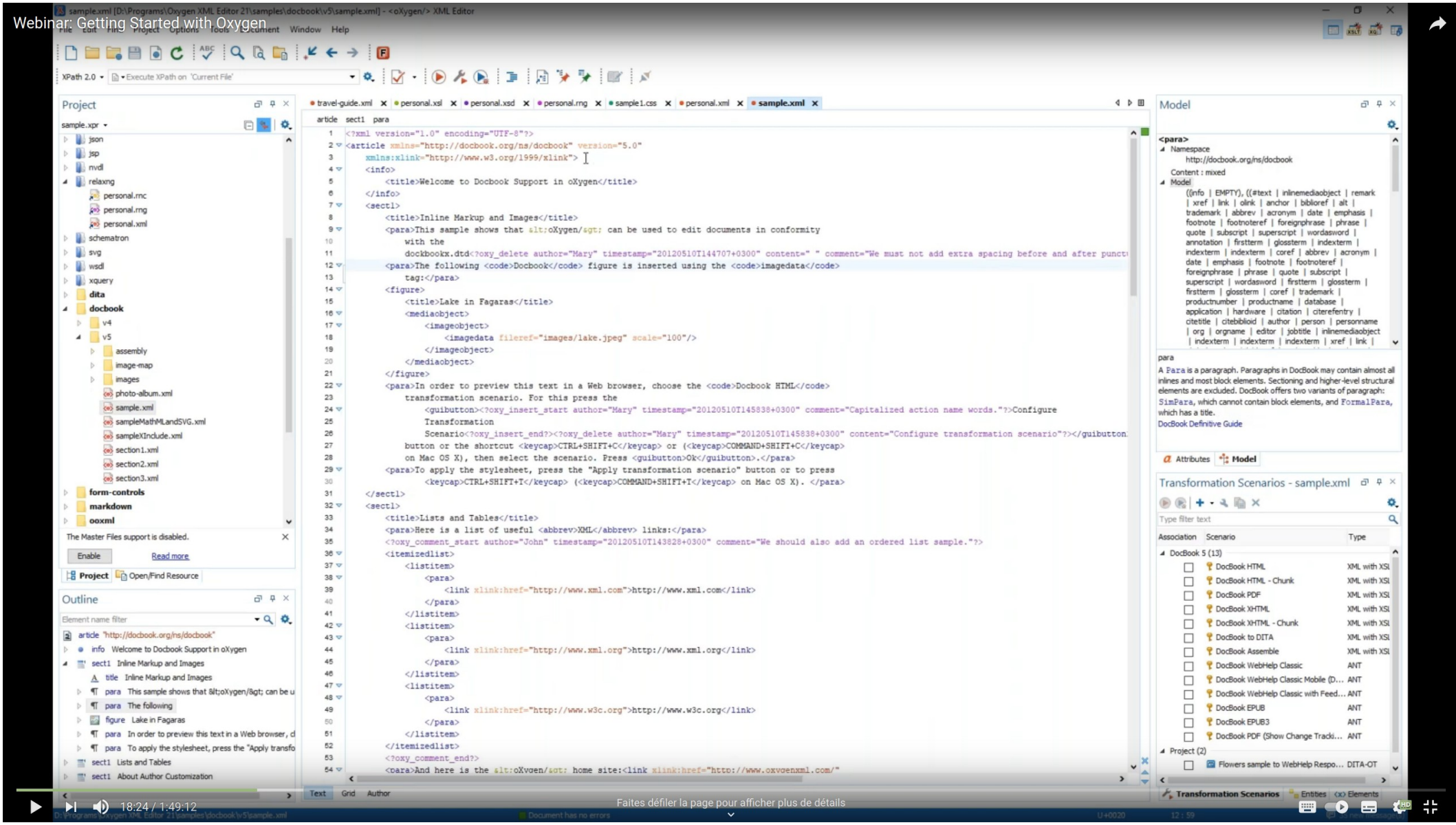
	User-based		Floating	
	License + 1-year SMP	License	License + 1-year SMP	License
Professional	$832 Add to cart	$688 Add to cart	$2,497 Add to cart	$2,064 Add to cart
Enterprise	$1,155 Add to cart	$955 Add to cart	$3,468 Add to cart	$2,866 Add to cart
Academic	$119 Add to cart	$99 Add to cart	$357 Add to cart	$297 Add to cart
Personal	$240 Add to cart	$198 Add to cart		

Oxygen XML Editor



https://www.youtube.com/watch?v=PiCWAIiKx78

Oxygen XML Editor



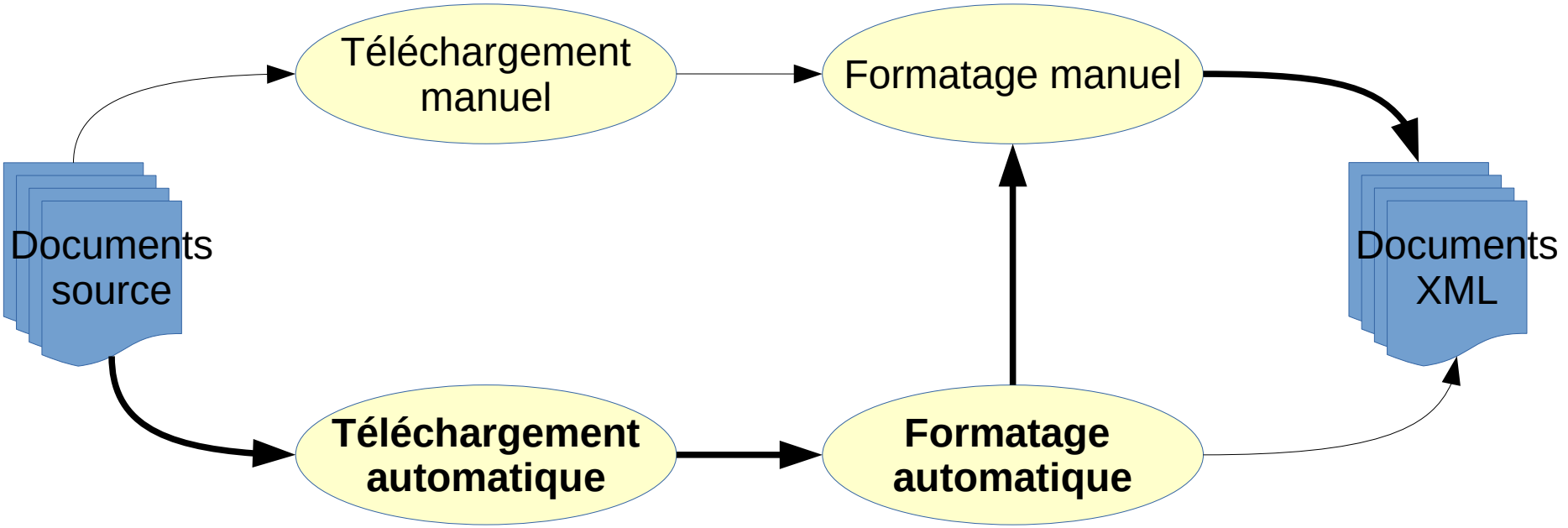
https://www.youtube.com/watch?v=PiCWAIiKx78

Difficulté de conversion vers XML

- XML, HTML (page Web)
  - Assez facile : souvent il suffit de renommer/supprimer des balises !
- Texte, DOC, ODT
  - Plus difficile, mais on peut souvent faire pas mal de choses avec du chercher/remplacer avancé (regex, formatage conditionnel...)
- PDF
  - Difficile à très difficile : souvent inexploitable
- De manière générale
  - Le travail est généralement différent pour chaque *source* de documents, donc on tend à réduire le nombre de sources.

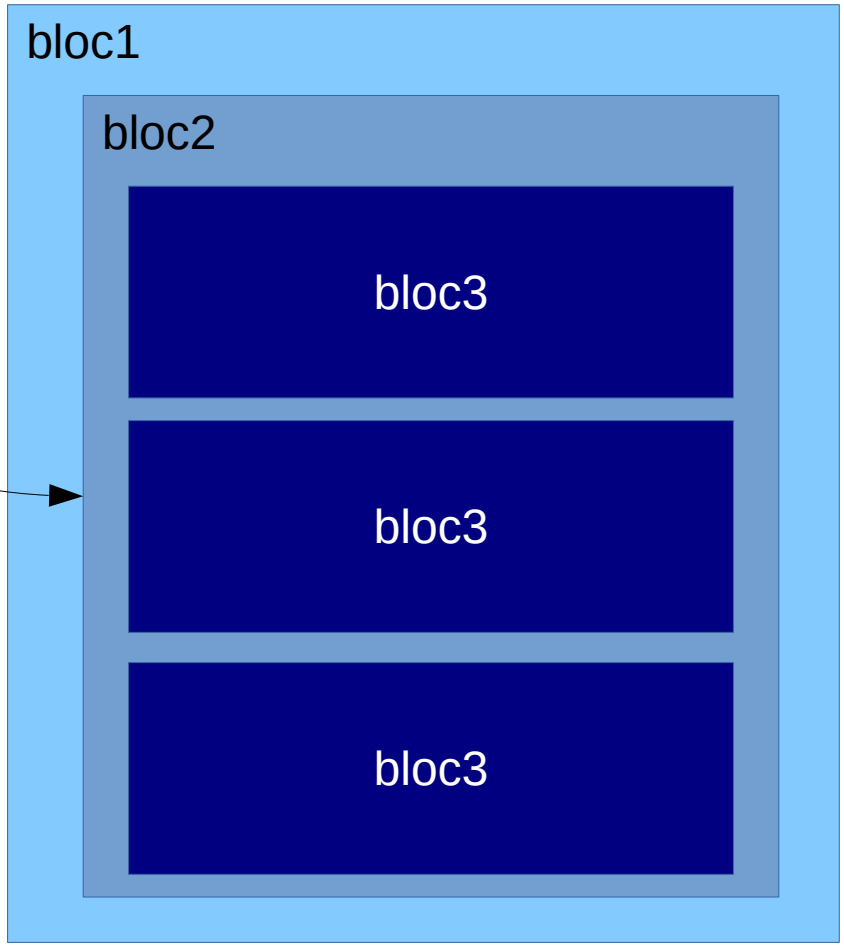
Collecte de documents Web

- Manuel / Automatisé



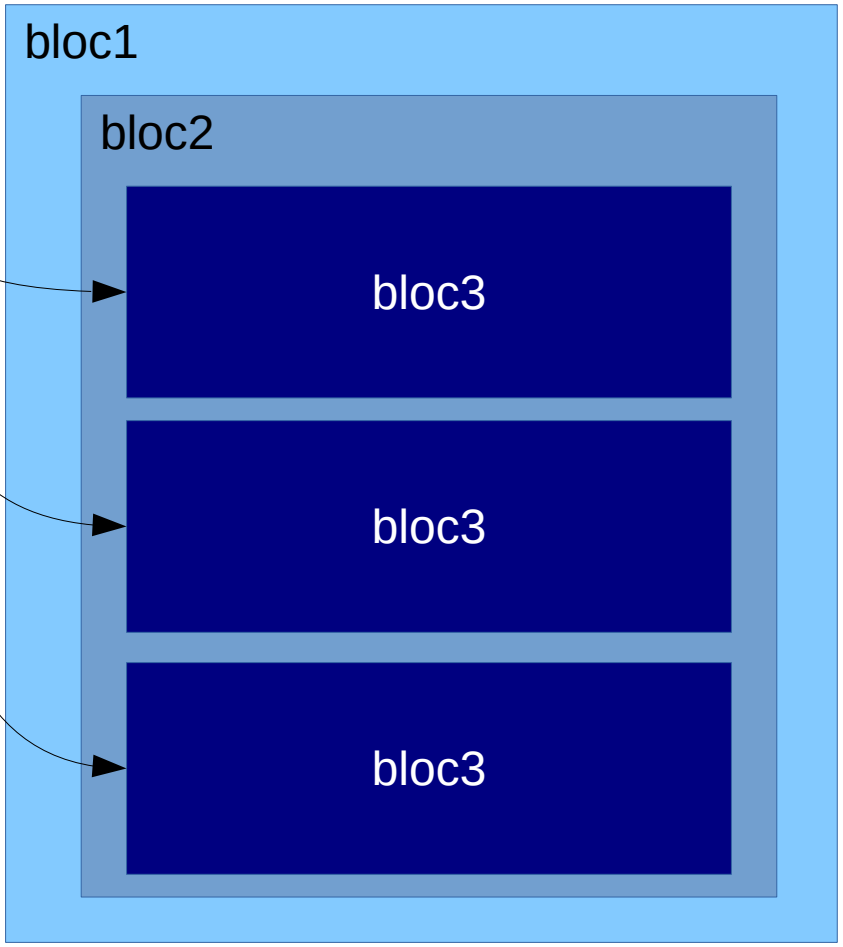
Le langage XPath

- Pour manipuler du XML
  - /bloc1/bloc2



Le langage XPath

- Pour manipuler du XML
  - /bloc1/bloc2/bloc3

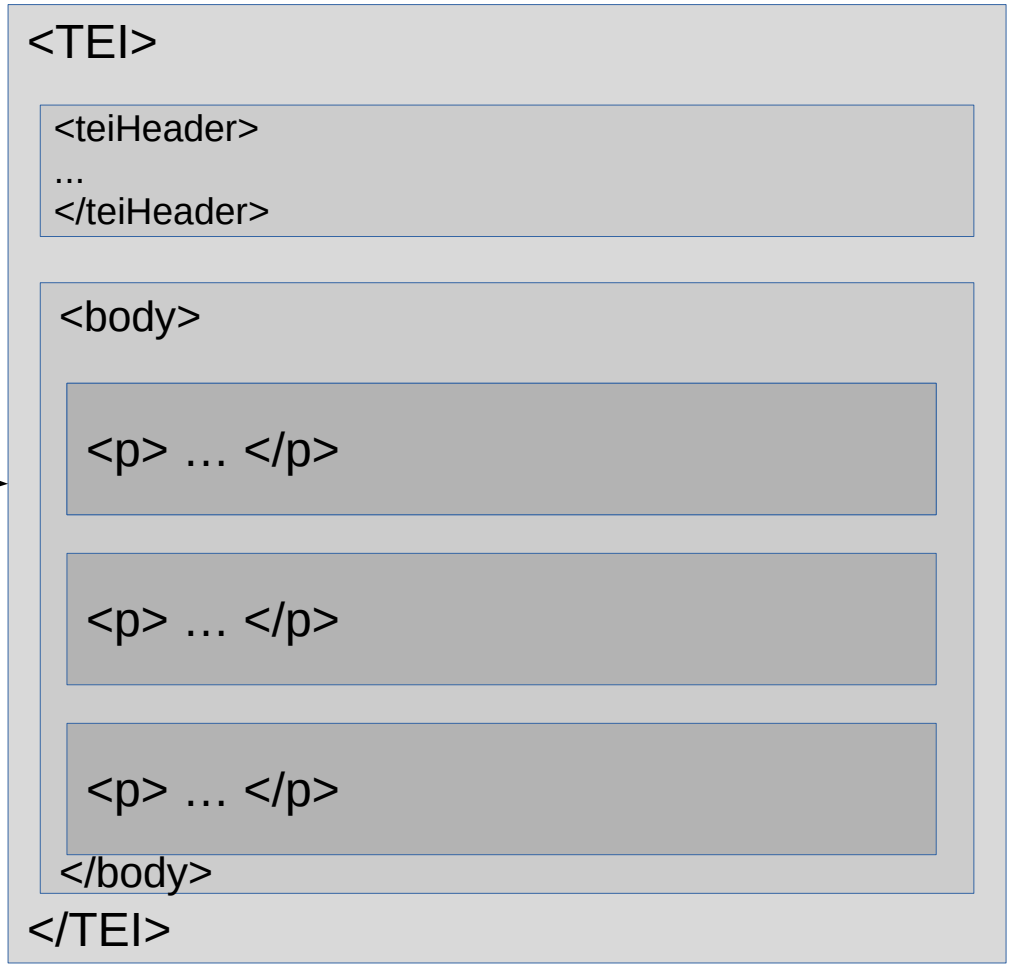


Le langage XPath

- Avec un vrai document
  - Test avec :
    - le fichier XML
      https://pro.aiakide.net/cours/Corpus2021a/exempleCorpus14.xml
    - le site d’évaluation d’expressions Xpath
      http://xpather.com/

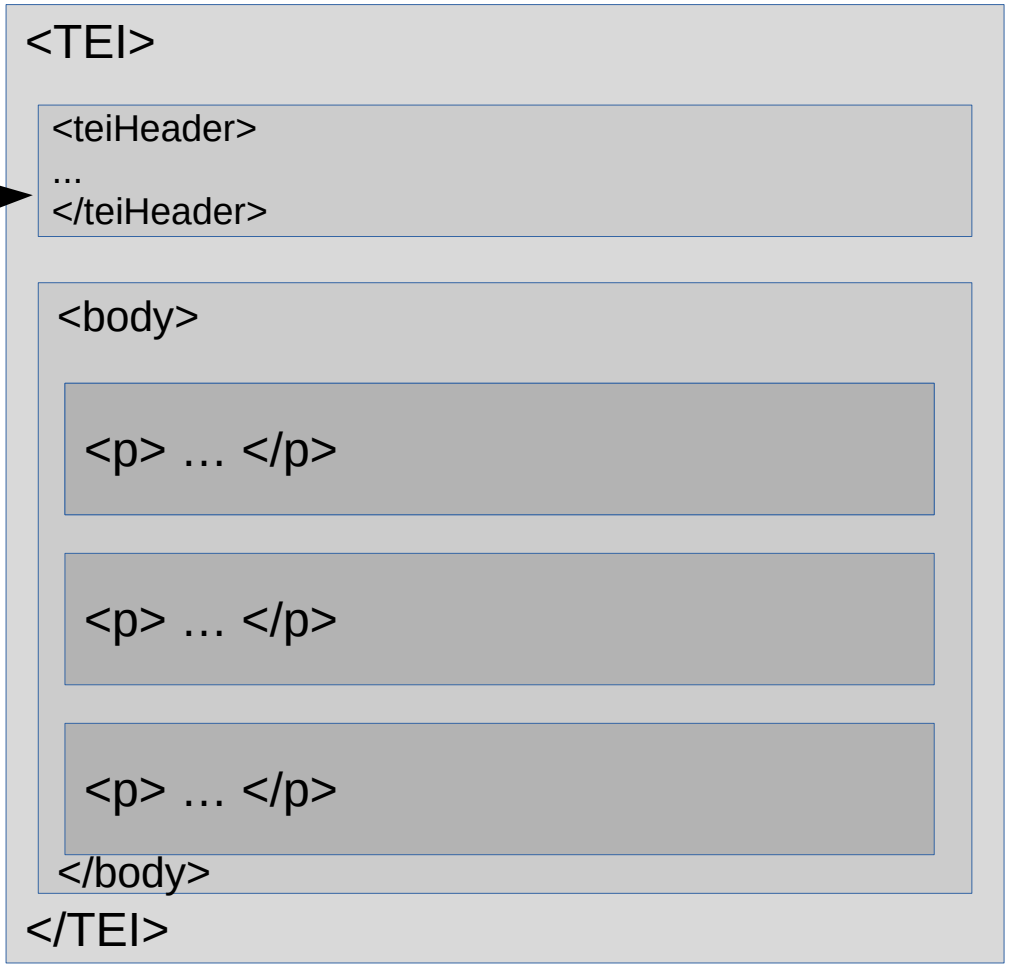
Racine du document

- Attraper tout le document :
  - /TEI



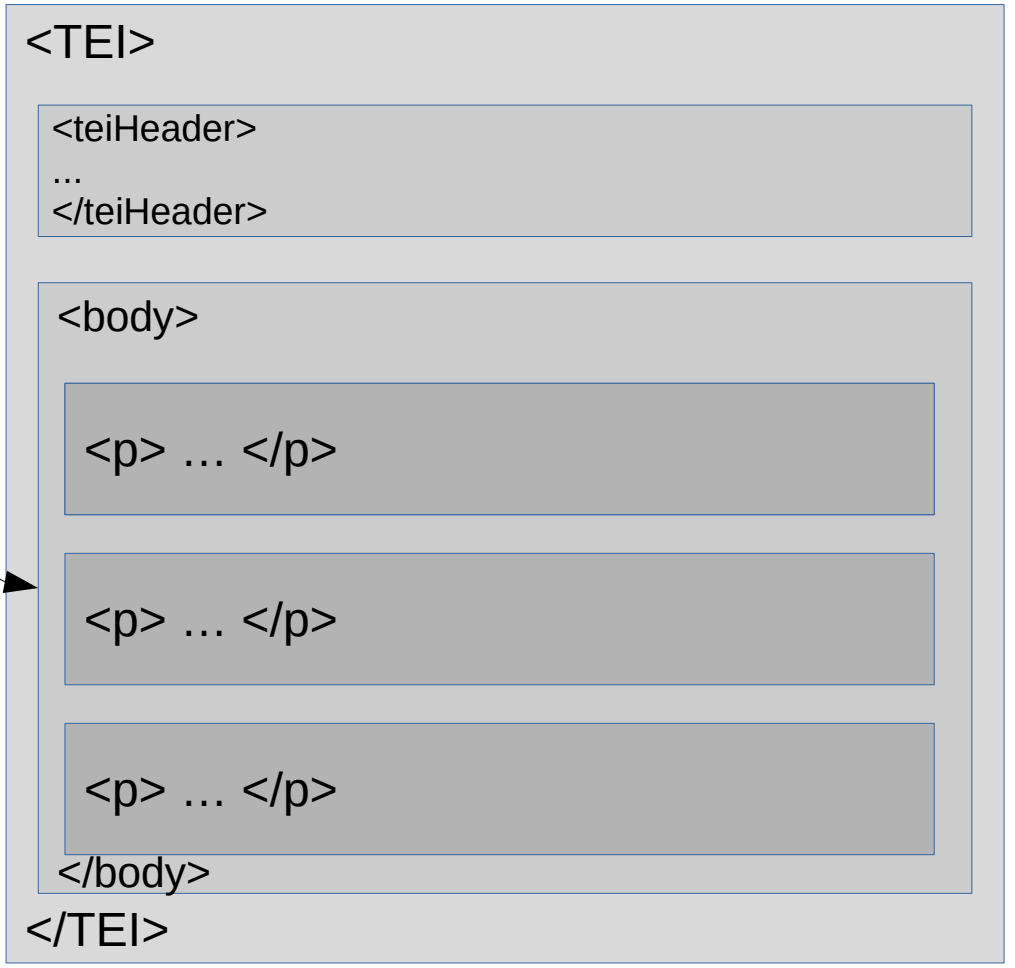
Un sous-bloc

- Attraper tout les blocs *teiHeader*:
  - /TEI/teiHeader



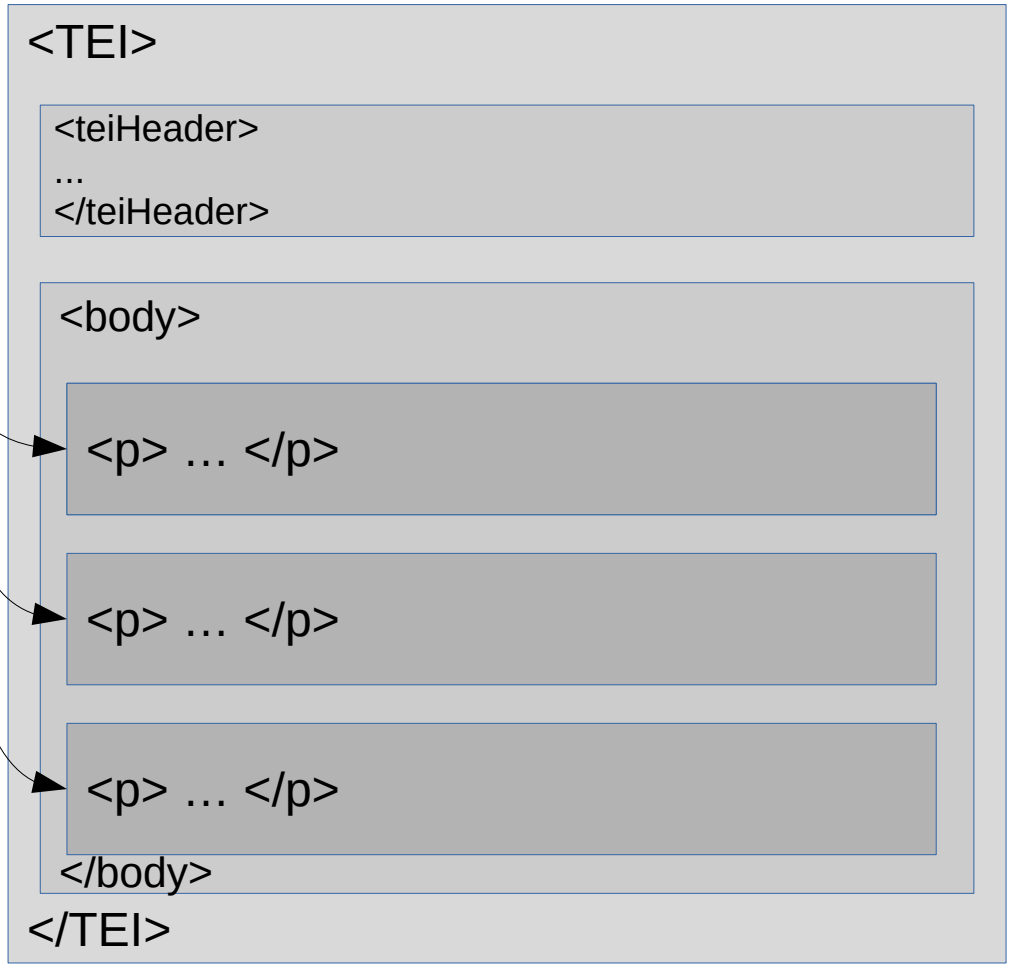
Un autre sous-bloc

- Attraper tous les blocs *body*:
  - /TEI/body



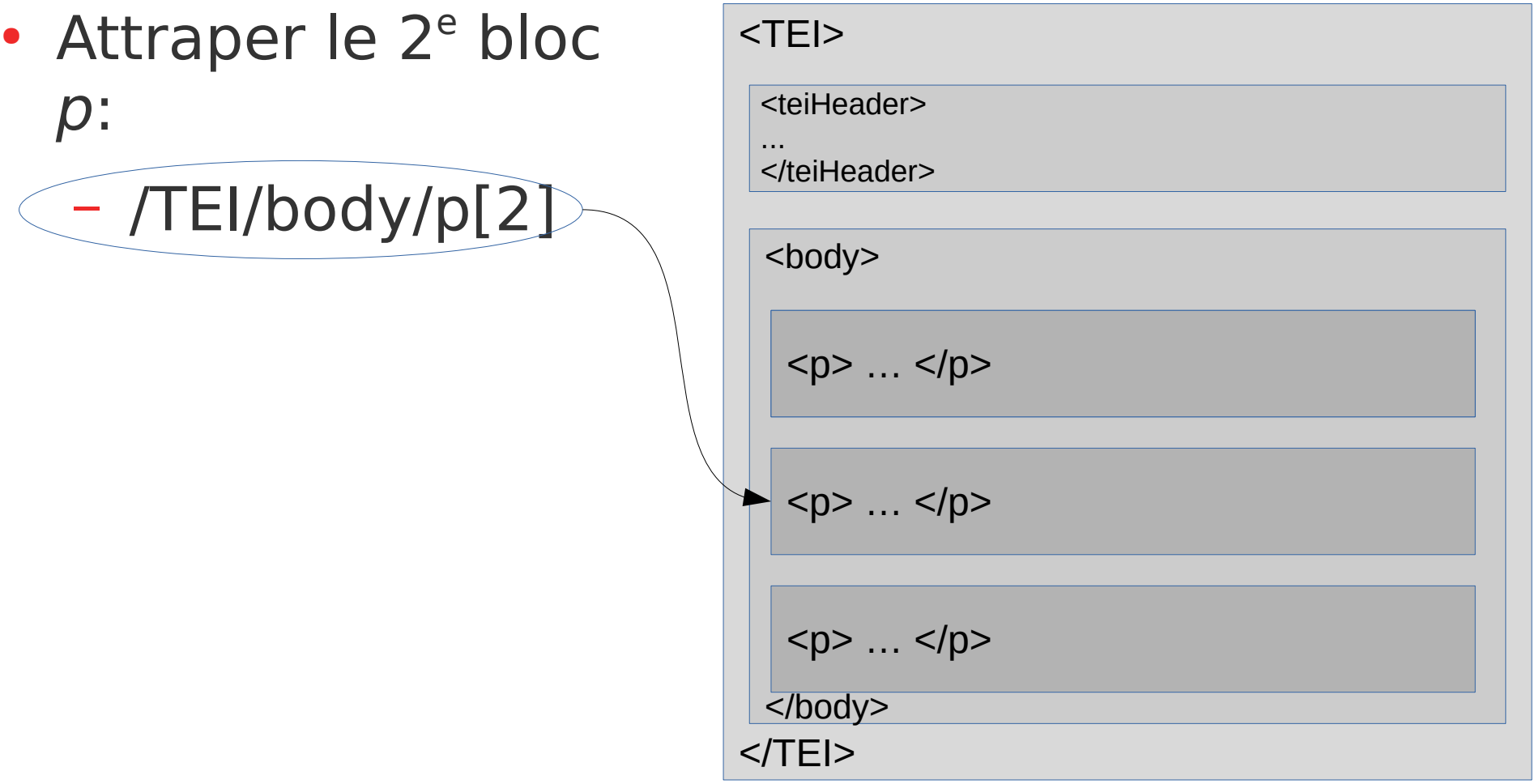
Un sous-sous-bloc

- Attraper tout les blocs *p*:
  - /TEI/body/p



Un sous-sous-bloc n°*x*

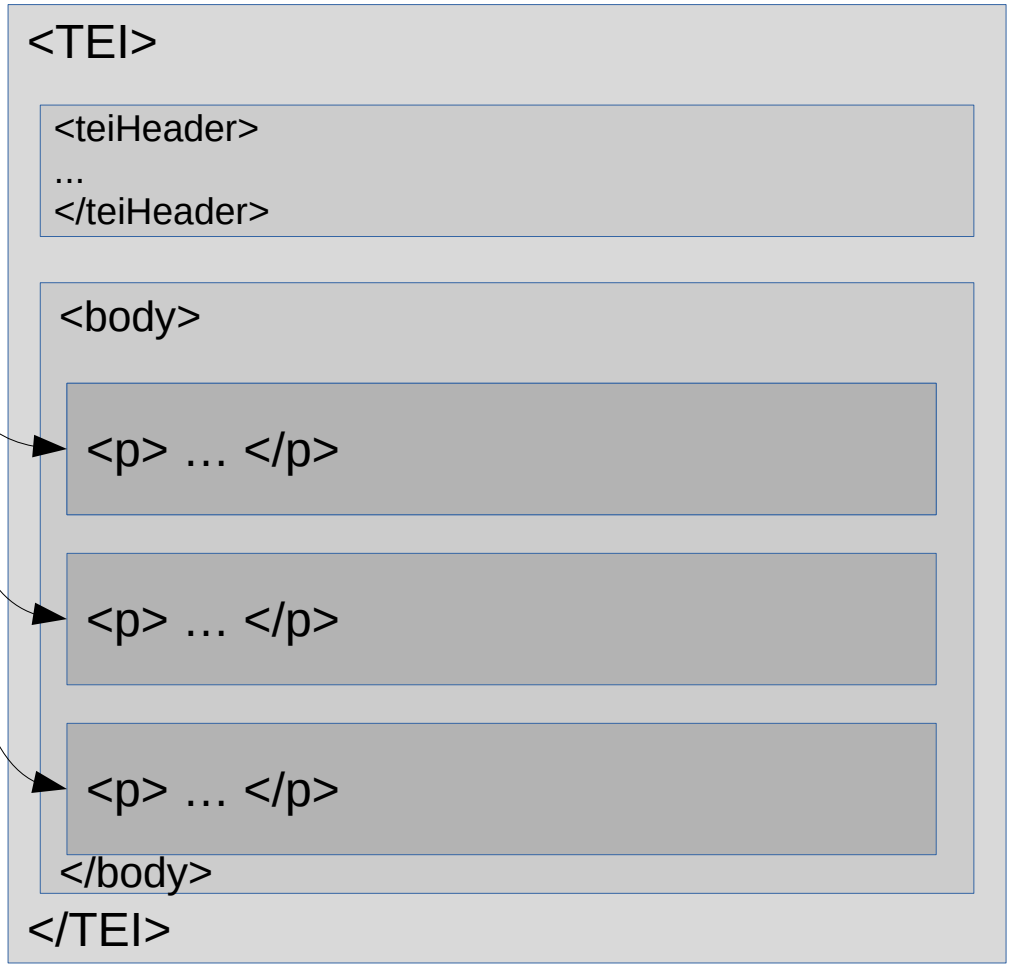
- Attraper le 2e bloc *p*:
  - /TEI/body/p[2]



Tous les blocs d’un type donné

- Attraper tous les blocs *p*:
  - //p

Où qu’ils soient dans le document !



Tous les blocs d’un type donné

- Attraper tous les blocs *availability*:
  - //availability

Tous les blocs d’un type donné… avec un attribut donné

- Attraper tous les blocs *availability* dont l’attribut *status* vaut *restricted*:
  - //availability[@status="restricted"]
- Attraper tous les blocs *p* dont l’attribut *xml:lang* vaut *fra*:
  - //p[@xml:lang="fra"]

Exercice

- À partir du document
  - https://pro.aiakide.net/cours/Corpus2021a/laurent-1-150216-2.xml
- Trouver…
  - tous les paragraphes
  - tous les paragraphes du texte
  - le 3e paragraphe du texte
  - tous les passages ajoutés par l’éditeur (bloc *add*)
  - tous les passages raturés (bloc *del* avec attribut rend="overstrike")

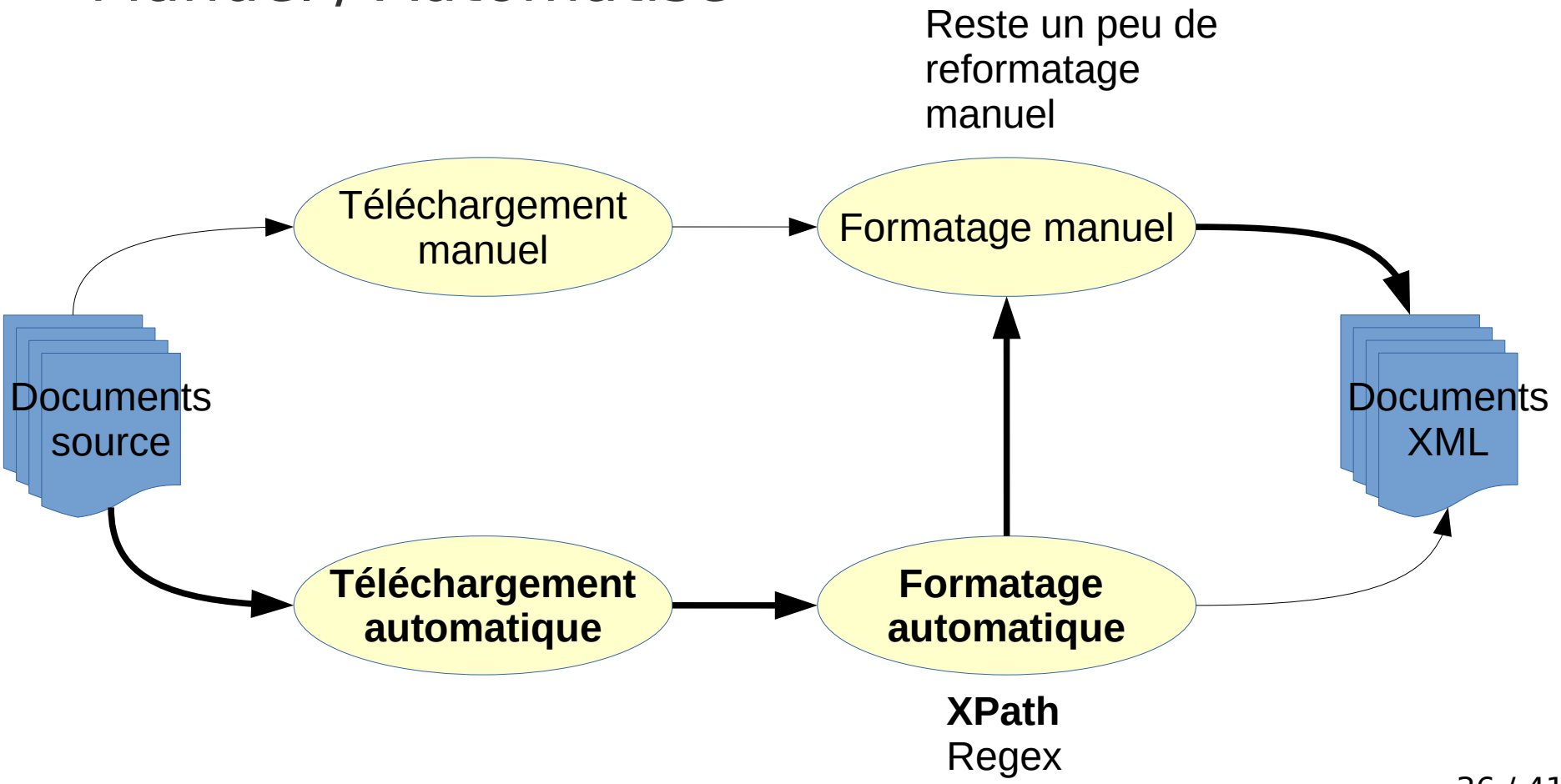
Utilisation en *scraping*

- Pour *scraper* du HTML
  - Les pages Web (HTML) sont aussi des documents XML !
  - Exemple pour *scrapper* plusieurs articles du journal *Le Monde*
    - Avec un outil Web : http://www.urlitor.com/web-scraping
    - Avec un outil en ligne de commande : Xidel
      - xidel *http://maPage/web* -e '/le/chemin/xpath' --xml

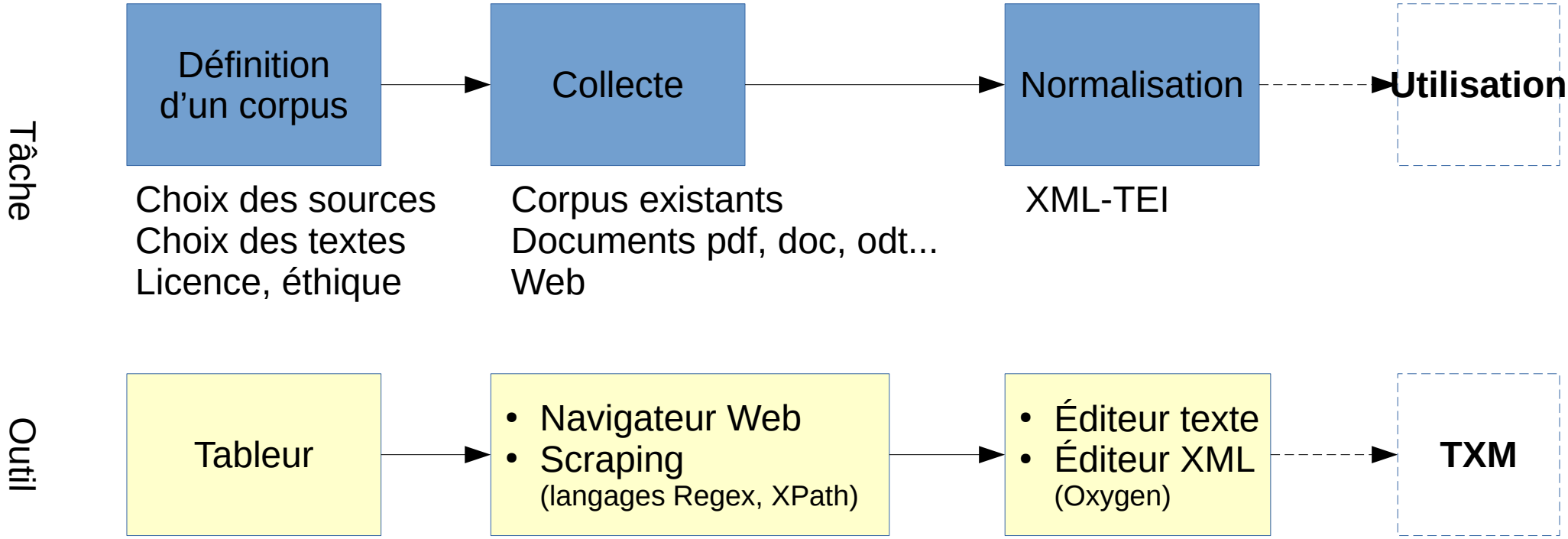

```
for url in $(cat listeURLs.txt) ; do xidel -e '/le/chemin/xpath' --xml ; done >> monCorpus.xml
```

Collecte de documents Web

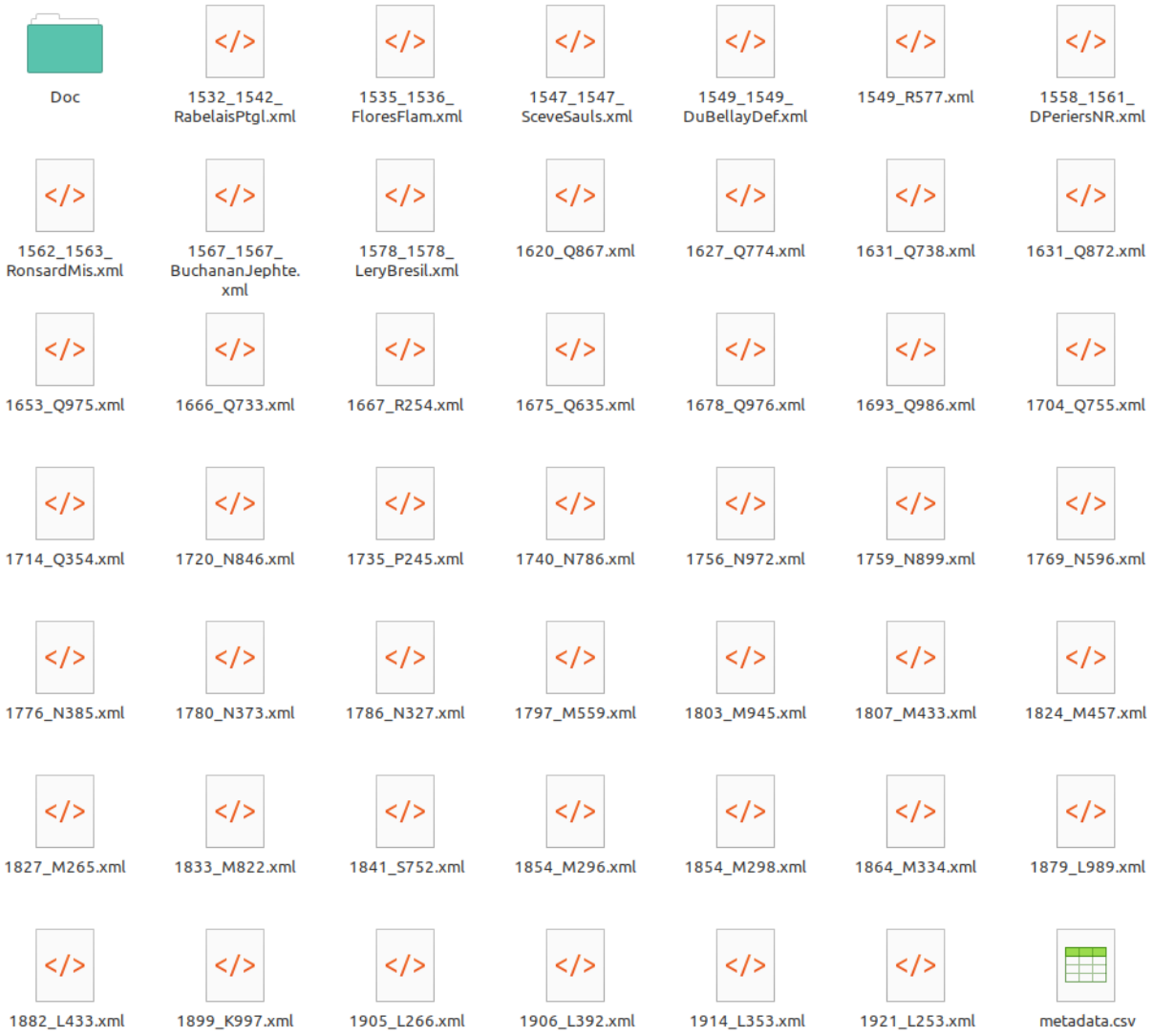
- Manuel / Automatisé



Où en sommes nous ?



Exemple : corpus Presto



Le fichier de métadonnées

metadata.csv

Métadonnées des textes

https://groupes.renater.fr/wiki/txm-users/public/tutoriel_import_txt_csv

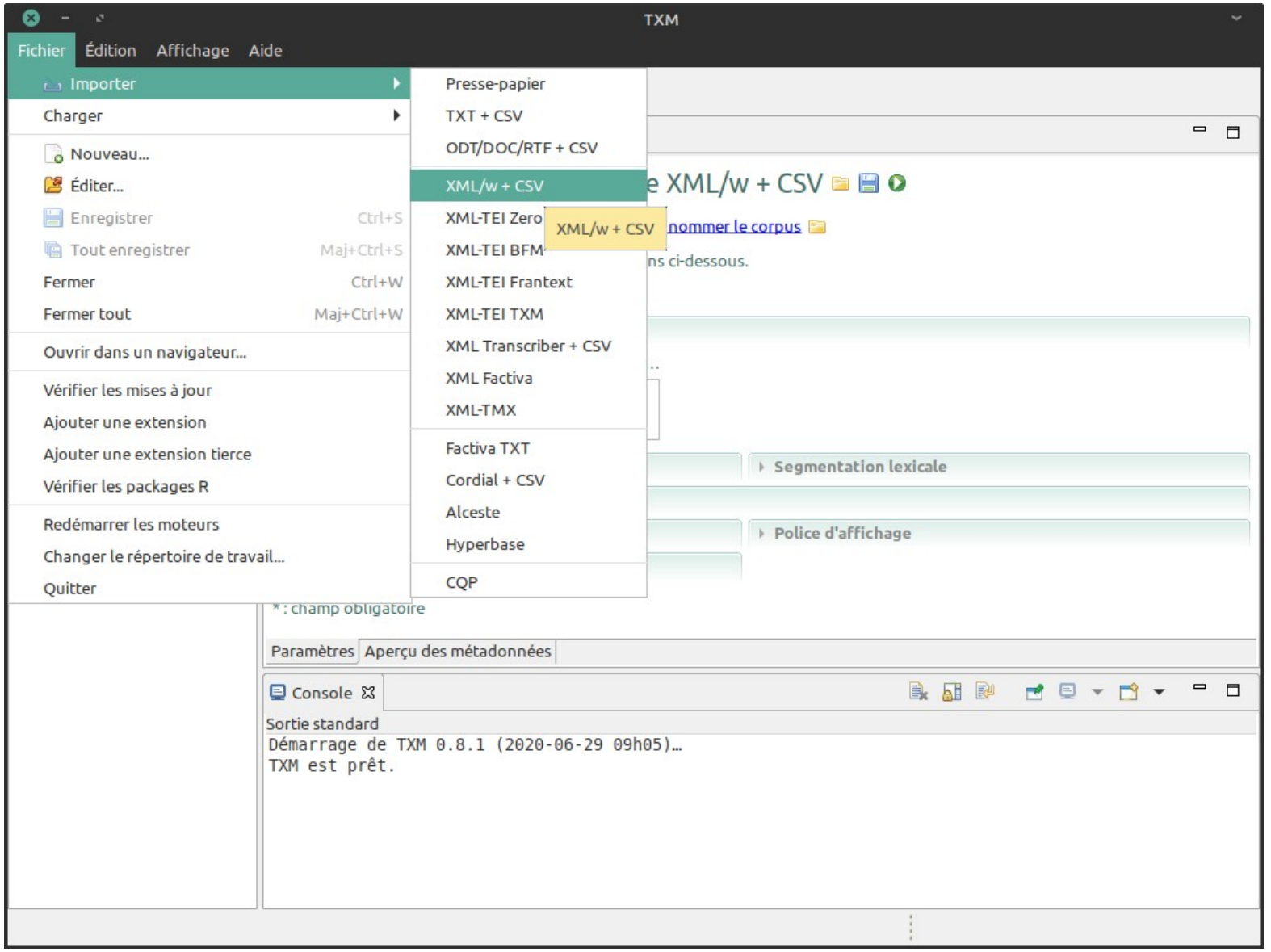
Pour associer des propriétés, appelées métadonnées, à chaque texte du corpus (par exemple, auteur, titre, date de production, genre, etc.), vous devez placer un fichier “metadata.csv” dans le même dossier que les textes à importer. Ce fichier est décrit dans la section 16.1 du Manuel de TXM (http://txm.sourceforge.net/doc/manual/manual59.xhtml#toc193).

Vous pouvez créer et éditer les métadonnées dans un logiciel tableur (Calc ou Excel, par exemple) et l'enregistrer au format CSV en utilisant les paramètres suivants (sous Excel, il faut sélectionner le format d'export “texte” et non “csv” afin de pouvoir contrôler ces paramètres) :

- le séparateur de colonne est virgule « , » ;
- le séparateur de texte est apostrophe droite double « “ » ;
- l'encodage des caractères doit être Unicode UTF-8 ;

	A	B	C	D	E	F	G	H	I	J
1	id	type	author	title	century	dateFirstEdition	dateCopyEdition	genre	theme	licence
2	1562_1563_RonsardMis	BVH-mod	RONSARD Pierre de	Discours des Miseres de ce temps	16	1562	1563	discours		Creative Commons BY-NC-SA 3.0
3	1582_S404	FRANTEXT	MONTAIGNE Michel de	Essais	16	1582	1595	traité	philosophie	Oeuvre sous droits
4	1578_1578_LeryBresil	BVH-mod	LÉRY Jean de	Histoire d'un voyage fait en la terre du Bresil	16	1578	1578	genre narratif		Creative Commons BY-NC-SA 3.0
5	1567_1567_BuchananJephte	BVH-mod	BUCHANAN George	Jephté, ou le veu	16	1567	1567	théâtre		Creative Commons BY-NC-SA 3.0
6	1549_1549_DuBellayDef	BVH-mod	DU BELLAY Joachim	La deffence, et illustration de la langue francoyse	16	1549	1549	traité	littérature	Creative Commons BY-NC-SA 3.0
7	1535_1536_FloresFlam	BVH-mod	FLORES Jean de	La Deplourable fin de Flamete	16	1535	1536	genre narratif		Creative Commons BY-NC-SA 3.0
8	1558_1561_DPeriersNR	BVH-mod	DES PÉRIERS Bonaventure	Nouvelles recreations et joyeux devis	16	1558	1561	genre narratif		Creative Commons BY-NC-SA 3.0
9	1532_1542_RabelaisPtgl	BVH-mod	RABELAIS François	Pantagruel	16	1532	1542	genre narratif		Creative Commons BY-NC-SA 3.0
10	1547_1547_SceveSauls	BVH-mod	SCÈVE Maurice	Saulsave	16	1547	1547	poésie		Creative Commons BY-NC-SA 3.0

Importation dans TXM



Pratique

- À partir du document
  - https://pro.aiakide.net/cours/Corpus2021a/laurent-1-150216-2.xml
- Trouver…
  - tous les paragraphes
  - tous les paragraphes du texte
  - le 3e paragraphe du texte
  - tous les passages ajoutés par l’éditeur (bloc *add*)
  - tous les passages raturés (bloc *del* avec attribut rend="overstrike")