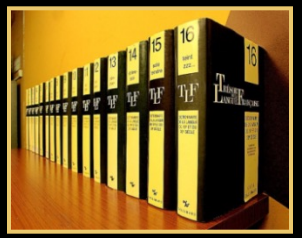
atilf
ANALYSE ET TRAITEMENT INFORMATIQUE DE LA LANGUE FRANÇAISE

TLFi
Trésor de la Langue Française Informatisée

14 décembre 2023
jessika.cardinali@atilf.fr






cnrs
UNIVERSITÉ DE LORRAINE

RAPPEL HISTORIQUE TLF

- **TLF :**

  dictionnaire des XIX^{e} et XX^{e} siècles en 16 volumes et 1 supplément (100.000 mots avec leur histoire, 270.000 définitions, 430.000 exemples…) paru entre 1971 et 1994.

  ***Avertissement** : la rédaction du TLF est terminée depuis 1994 et la plupart des contributeurs ont quitté le laboratoire. Il n'a pas vocation à être mis à jour. Cette ressource, qui ne fait pas l'objet d'une veille lexicographique, est donc close « en l'état ». Il est donc tout à fait naturel que les définitions qui s'y trouvent ne rendent pas compte des évolutions de la société.*

RAPPEL HISTORIQUE TLFi

- **TLFi :**
  - Début de la réflexion du projet d'informatisation du TLF en 1992 :

- Saisie des tomes 1 à 8 grâce à un accord CNRS-BNF :
  - Marché passé avec la société Berger-Levrault
  - Saisie faite en fonction d'une maquette
  - Livraison des tomes en 1997

- Les tomes 9 à 16 existent sous forme d'archives de photocomposition :
  - Rétro-conversion réalisée au laboratoire permettant le passage de l'état linéaire des bandes de photocomposition à un état structuré

- 81 fichiers XML et une DTD
- Mise en ligne en 2002 sur le site http://atilf.atilf.fr/
- Publication d'un CD-ROM en 2004
- Mise en ligne en 2005 sur le CNRTL

OBJECTIF DU PROJET

- Groupe de travail composé de Evelyne Jacquey, Sandrine Ollinger, Etienne Petitjean, Jessika Cardinali

- Objectif initial :

  Traduire le balisage du TLFi au format standard TEI et le rendre accessible via ORTOLANG, mais :

  - Le modèle dictionnaire de la TEI n'est pas adapté à la structure diversifiée et complexe du TLFi.

  - Cette constatation a également été faite par Jean Véronis lors d'un colloque en 1995.

- Objectif final :

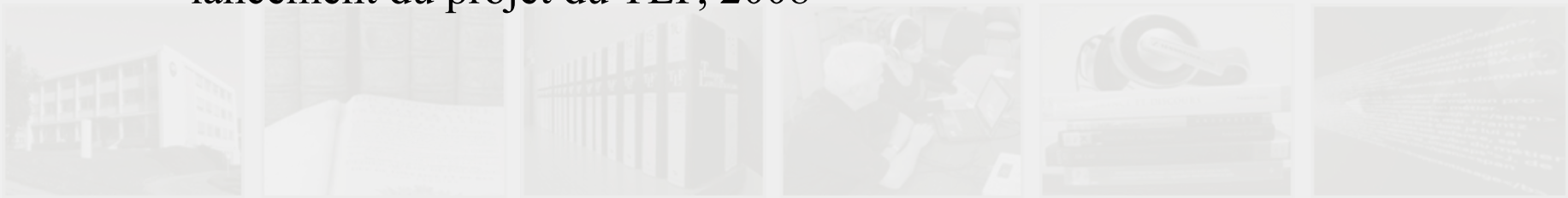
  Traduire le balisage du TLFi dans un autre modèle plus lisible avec une documentation pour :

  - fixer une version en UTF-8
  - le rendre accessible via ORTOLANG
  - être utilisée par de futurs groupes de travail

DOCUMENTS DE TRAVAIL

- Normes de rédaction (dossier prêté par Alain, un autre dossier a été découvert aux archives, à trier et numériser)

- Ouvrages trouvés au CDD :
  - *Informatique et Lexicographie : à propos de l'informatisation du dictionnaire*, Colette LUC, Mémoire de Maîtrise, 1992-1993
  - *Autour de l'informatisation du TLF*, Actes du Colloque Internationale de Nancy, 1995
  - Pré-Acte du Colloque Internationale à l'occasion du 50^{e} anniversaire du lancement du projet du TLF, 2008



ORGANISATION DU TRAVAIL

- QUAND ?
  - Début du groupe de travail : 11 janvier 2022 (première réunion)
- COMMENT ?
  - GIT : 245 tickets créés
  - Réunion hebdomadaire (mardi après-midi)
- POURQUOI ?
  - Comprendre la structuration du balisage : 76 balises différentes
  - Réaliser la chaine de traitement jusqu'à la version TLFi 2023

ÉTAPE DU TRAVAIL : DÉFINIR LA SOURCE PRIMAIRE

Plusieurs sources existent :

- Fichiers issus du CNRTL (pas d'historique de cette version)
- Fichiers de 2011 (travail d'amélioration de la hiérarchie de la balise <H> fait à partir des fichiers issus du CNRTL)

Une comparaison a été faite et des différences de structure et de contenu ont été relevés.

Décision prise :

- Utilisation de la source de 2011 qui contient les corrections de structure
- Étude des différences de contenu et apporter des modifications si nécessaire (travail fait par Sandrine, en lien avec Alain et Yan)

Structure d'une entrée

- Nombre d'entrées : plus de 54000

- Information sur la vedette
- Partie synchronie
- Différentes rubriques

```
<art id="11012">
    <ved>
        <mot>
            <da>
                <R>BADINAGE</R>
            </da>
        </mot>
        <separateur>
            <da>
                <R>, </R>
            </da>
        </separateur>
        <cod>
            <da>
                <R>subst. masc.</R>
            </da>
        </cod>
    </ved>
    <sync> [641 lines]
    <rpro> [5 lines]
    <rety> [35 lines]
    <rsta> [5 lines]
    <rbbg> [8 lines]
</art>
```

ÉTAPE DE TRAVAIL : ÉTUDE DU BALISAGE

- Étude de chaque balise :
  - Sa fréquence
  - Son environnement → Choix final
  - Son contenu

- Soucis rencontrés :
  - Contenu du TLFi différent du TLF : corrections manuelles du contenu (par exemple des caractères mal rétro-convertis), mauvaise gestion des espaces (à cause des indentations précédentes ?)
  - Mauvais balisage du contenu : le contenu textuel ne correspond pas au nom de la balise (par exemple un titre peut être balisé <auteur>), texte mis en commentaire, mauvais balisage des rubriques

CRÉATION D'UNE SOURCE INTERMÉDIAIRE : Modifications

- Décision de créer une source intermédiaire avec pour modifications :
  - Conversion des balises contenant des caractères en unicode : oe ou OE, caractères grecs, caractères phonétiques… (<car v="oe"/>)
  - Beaucoup de commentaires avec pour contenu des signes de ponctuation (<!--<R>).</R>-->) -> décommenter et introduction d'une nouvelle balise : <separateur>
  - Beaucoup de balises vides : suppression
  - Rebalisage des données textuelles (regroupement de toutes les données textuelles dans un seul <da> :

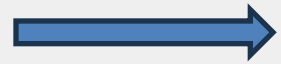
```
<da>
    <R>13. ... M</R>
</da>
<EXP>
    <da>
        <R>me</R>
    </da>
</EXP>
<da>
    <R>Aubry, partenaire habitu
    <G>abordages</G>
    <R> tumultueux.
    </R>
</da>
```

```
<da>
    <R>13. ... M</R>
<EXP>
    <da>
        <R>me</R>
    </da>
</EXP>
    <R>Aubry, partenaire habi
    <G>abordages</G>
    <R> tumultueux.
    </R>
</da>
```

CRÉATION D'UNE SOURCE INTERMÉDIAIRE : Modifications

- Modification _ en cadratin
- Déplacement de contenu :
  - Déplacement d'une virgule ou d'un point se trouvant à la fin de <mot> dans une balise <separateur>

```
<mot>
   <da>
      <R>A,</R>
   </da>
</mot>
```



```
<mot>
   <da>
      <R>A</R>
   </da>
</mot>
<separateur>
   <da>
      <R>, </R>
   </da>
</separateur>
```

CRÉATION D'UNE SOURCE INTERMÉDIAIRE : Modifications

- Déplacement d'une parenthèse se trouvant devant une source bibliographique

```
<def n="e">
    <da>
        <R>,,S'approcher de l'endroit où la perdrix s'est réfugiée.`` (</R>
    </da>
    <so>
        <da>
            <I>Ac.</I>
            <R> 1842</R>
        </da>
    </so>
    <!--<R>).</R>-->
</def>
```

```
<def n="e">
    <da>
        <R>,,S'approcher de l'endroit où la perdrix s'est réfugiée.`` </R>
    </da>
    <separateur>
        <da>
            <R>(</R>
        </da>
    </separateur>
    <so>
        <da>
            <I>Ac.</I>
            <R> 1842</R>
        </da>
    </so>
    <separateur>
        <da>
            <R>).</R>
        </da>
    </separateur>
</def>
```

CRÉATION D'UNE SOURCE INTERMÉDIAIRE : Gestion des espaces

- Beaucoup d'espaces sont manquants :
  - Après <EXP> et <IND> (plus de 170 000) : étude faite du contenu de <EXP> et <IND> et de ce qui suit pour décider

```
<da>
   <R>13. ... M</R>
</da>
<EXP>
   <da>
      <R>me</R>
   </da>
</EXP>
<da>
   <R>Aubry, partenaire habituel
   <G>abordages</G>
```

```
<da>
    <R>13. ... M</R>
    <EXP rendu="R">me</EXP>
    <R> Aubry, partenaire ha
    <G>abordages</G>
```

  - Ajout d'espaces avant ou après des signes de ponctuation

CRÉATION D'UNE SOURCE INTERMÉDIAIRE : Restructuration des rubriques

- Constatation d'un mauvais balisage des rubriques

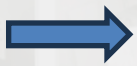
  Exemple pour la rubrique étymologie : seulement 8010 occurrences dans la source primaire pour plus de 54000 entrées

- Études des 11 rubriques différentes :

  Par quoi commence chaque rubrique : régularité <G>xxx</G>

- Modification du balisage avec un script python

```
<rpro>
   <da>
      <G>Prononc. :</G>
      <R> [badmintɔn]. </R>
      <I>Lar. encyclop.</I>
      <R> donne la transcr. : badminton. </R>
      <G>Étymol. et Hist.</G>
      <R> 1898 jeux (</R>
      <C>G. de Saint Clair</C>
```



```
<rpro>
    <da>
        <G>Prononc. :</G>
        <R> [badmintɔn]. </R>
        <I>Lar. encyclop.</I>
        <R> donne la transcr. : badminton. </R>
    </da>
</rpro>
<rety>
    <da>
        <G>Étymol. et Hist.</G>
        <R> 1898 jeux (</R>
        <C>G. de Saint Clair</C>
```

Traitement des codes grammaticaux

Certains codes sont constitués d'un code grammatical et de texte autre que de l'information grammaticale

- Restructuration de ces codes et introduction d'une balise <cod2> dans la source primaire

```
<cod>
   <da>
      <R>subst. masc. et adj. (On dit aussi </R>
      <I>Abdérite</I>
      <R>).</R>
   </da>
</cod>
<cod2>
   <posCod>
      <da>
         <R>subst. masc. et adj.</R>
      </da>
   </posCod>
   <separateur>
      <espace/>
   </separateur>
   <precision type="ind">
      <da>
         <R>(On dit aussi </R>
         <I>Abdérite</I>
         <R>)</R>
      </da>
   </precision>
   <separateur>
      <da>
         <R>. </R>
      </da>
   </separateur>
</cod2>
```

Traitement des codes grammaticaux

Certains mots ont une balise <cod> ou <RCOD> contenant seulement un tiret. Ce tiret n'est pas présent dans la version TLF. On suppose que ce tiret a été introduit lors de la rétro-conversion.

- Introduction d'une balise <cod2> vide dans la source primaire

```
<cod>
   <da>
      <R>-</R>
   </da>
</cod>
<cod2>
   <posCod>
      <da>
         <R/>
      </da>
   </posCod>
</cod2>
```

- Suppression du tiret et transformation en <cod/> dans la source intermédiaire

```
<cod/>
<cod2>
    <posCod>
        <da>
            <R/>
        </da>
    </posCod>
</cod2>
```

Traitement des codes grammaticaux

Pour les autres <cod> non problématiques : introduction d'une balise <cod2> en copiant le contenu de <cod>

```
<cod>
    <da>
        <R>subst. masc.</R>
    </da>
</cod>
<cod2>
    <posCod>
        <da>
            <R>subst. masc.</R>
        </da>
    </posCod>
</cod2>
```

Traitement des codes grammaticaux : import et expansion

- <cod2> est vide : Importation du code grammatical avec un fichier csv (listing "entrée + code") et un script python

```
<cod2>
   <posCod>
      <da>
         <I/>
      </da>
   </posCod>
</cod2>
```

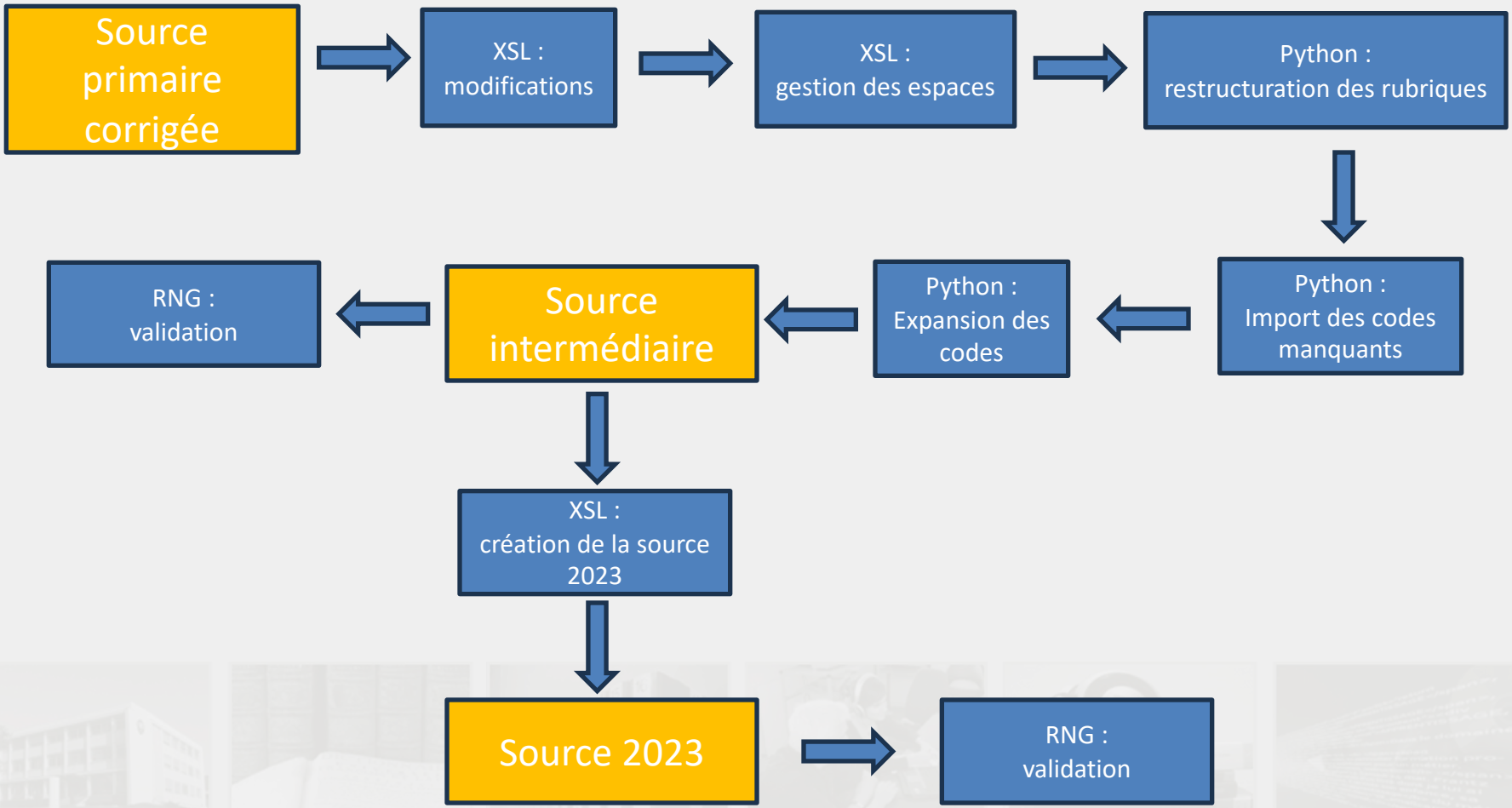
```
<cod2>
   <posCod>
      <da>
         <I>verbe</I>
      </da>
   </posCod>
</cod2>
```

- Expansion du code grammatical de chaque <cod2> avec un script python

```
<cod>
    <da>
        <R>subst. masc.</R>
    </da>
</cod>
<cod2>
    <posCod>
        <da>
            <R>subst. masc.</R>
        </da>
    </posCod>
</cod2>
```

```
<cod>
    <da>
        <R>subst. masc.</R>
    </da>
</cod>
<cod2>
    <posCod>
        <da>
            <R>substantif masculin</R>
        </da>
    </posCod>
</cod2>
```

Chaine de traitement



En attente

- Faire un point sur les espaces manquantes
- Typage plus fin des rubriques
- Des caractères présents n'existent pas en unicode, il faudra créer les images
- Corriger la bibliographie grâce à Frantext
- Rédiger la documentation

Mise à disposition de la ressource début 2024