Tout document TEI doit porter des métadonnées qui renseignent sur la ressource TEI elle-même et le texte qu'elle représente. Si le texte est la transcription d'un document source numérisé, qui est le cas pour tout document du projet *Gallic(orpor)a*, trois objets sont concernés. En plus de la ressource créée, il y a le document source physique qui a été numérisé ainsi que la numérisation elle-même qui a été traitée par les modèles HTR. Pour le projet *Gallic(orpor)a* qui avait pour but le traitement automatisé des fac-similés numériques, les métadonnées du document TEI portent sur les trois objets de texte suivants :

1. la ressource lexicographique numérique, c'est-à-dire le document TEI lui-même que le pipeline *Gallic(orpor)a* produit
2. la numérisation du document source, c'est-à-dire le fac-similé numérique stocké dans la base de données Gallica
3. le document source physique, qui appartient du fonds de la Bibliothèque nationale de France et a été numérisé

Ce chapitre explique de quoi consistent les métadonnées de ces trois objets de texte ainsi que le format par lequel elles sont accédées. Malgré la diversité de la portée du projet *Gallic(orpor)a*, chaque document traité concerne ces trois objets puisque chacun est un fac-similé numérique, dérivé d'un document source physique et transformé en ressource lexicographique par le pipeline. Pour être faites à l'échelle, les métadonnées doivent être les mêmes pour tout type de document. Un manuscrit écrit par plusieurs mains et apparu sans éditeur, issue d'un scriptorium à une date approximative, doit se disposer de mêmes types de métadonnées qu'un imprimé écrite par une autrice et publiée par un éditeur. Une solution pour surmonter ce défi est de minimiser les métadonnées portant sur le document source et d'encoder uniquement *l'essentiel*. Au sein de cette solution est la sélection d'en quoi constitue l'essentiel.

0.0.1 La structure du `<teiHeader>`

Les métadonnées de la ressource lexicographique du pipeline *Gallic(orpor)a* sont toutes encodées dans un élément XML que le schème TEI exige : le `<teiHeader>`. Dans le cadre du projet *Gallic(orpor)a*, le `<teiHeader>` servait à enrichir l'exploitation de la ressource numérique ainsi qu'à faciliter l'échange de ses données. Selon les normes de la TEI, trois parties peuvent constituer le `<teiHeader>` : une description bibliographique de l'encodage (`<fileDesc>`), une description des aspects non bibliographiques tel qu'un classement du contenu textuel qui appartiennent au texte représenté (`<profileDesc>`), et une description technique de l'encodage et/ou l'appareil qui l'a fait (`<encodingDesc>`).

1 La description bibliographique (`<fileDesc>`)

La première chose sur laquelle la ressource numérique doit renseigner est elle-même. Ces informations sont organisées dans l'élément `<fileDesc>`. Traduit littéralement en français comme *la description du fichier*, le `<fileDesc>` décrit

le document TEI lui-même. Au moins, cette description doit porter sur les trois aspects suivants :

1. le titre et la responsabilité de la ressource numérique (`<titleStmt>`)
2. la distribution de la ressource, y compris les droits d'utilisation (`<publicationStmt>`)
3. le document source dont le texte la ressource numérique représente (`<sourceDesc>`)

Un document TEI peut présenter d'autres éléments dans le `<teiHeader>` afin d'apporter encore plus de détail bibliographique. Par exemple, le `<editionStmt>` précise l'édition de l'œuvre que la ressource représente. Les documents TEI produits par le pipeline *Gallic(orpor)a* ne profitent pas de cet élément parce que certains documents du corpus traité, tel que les manuscrits, n'ont pas d'édition et on veut que chaque document TEI ait les mêmes types de métadonnées dans le `<teiHeader>`. Au contraire, d'autres éléments facultatifs, tel que le `<extent>`, sont produits par le pipeline. L'élément `<extent>` est utile parce qu'il porte sur la taille de la ressource, tel comme le nombre de pages transcrites qu'elle compte. La ressource lexicographique n'est pas toujours une transcription du document complet. Il faut préciser la quantité des pages transcrites.

1.1 Le titre et la responsabilité (`<titleStmt>`)

Le titre et la responsabilité de la ressource numérique sont tous les deux représentés dans l'élément `<titleStmt>`. Cet élément descendant du `<fileDesc>` peut avoir plusieurs descendants mais il faut au moins un titre. Après le titre (`<title>`) il est recommandé d'indiquer l'individu ou les individus qui sont responsables pour la création du texte représenté et/ou de la ressource numérique. Dans le cadre du projet *Gallic(orpor)a*, nous avons conçu un schème `<titleStmt>` simple qui porte sur le titre, l'auteur ou les auteurs du texte, et les personnes du projet responsables pour la création de la ressource numérique.

1.1.1 La responsabilité

Entre les lignes 4 et 21 de la Figure 1 on voit l'élément `<respStmt>` qui contient des informations sur l'équipe du projet *Gallic(orpor)a*. Tout document encodé par le pipeline s'informe sur le titre (`<title>`), la personne ou les personnes auxquels est attribuée la responsabilité du texte (`<author>`), et la responsabilité de notre équipe qui a conçu le pipeline et l'application `alto2tei` (`<respStmt>`). La déclaration de responsabilité peut être personnalisée selon le projet ou selon l'équipe qui utilise le pipeline *Gallic(orpor)a* ou l'application `alto2tei` pour créer la ressource numérique. En général, elle devrait contenir une phrase qui résume la nature de la création de la ressource, telle que la phrase « Transformation from ALTO4 to TEI by », balisée dans l'élément `<resp>`. Ensuite, elle devrait contenir des éléments `<persName>` qui renseignent sur les individus responsables, surtout en indiquant leurs noms.

```
1 <titleStmt>
2   <title>Titre du document source traité</title>
```

```
  <author>Auteur</author>
  <respStmt>
    <resp>Transformation from ALTO4 to TEI by</resp>
    <persName>
      <forename>Kelly</forename>
      <surname>Christensen</surname>
      <ptr type="orcid" target="000000027236874X"/>
    </persName>
    <persName>
      <forename>Simon</forename>
      <surname>Gabay</surname>
      <ptr type="orcid" target="0000000190944475"/>
    </persName>
    <persName>
      <forename>Ariane</forename>
      <surname>Pinche</surname>
      <ptr type="orcid" target="0000000278435050"/>
    </persName>
  </respStmt>
</titleStmt>
```

FIGURE 1 – Les informations sur le titre de la ressource

1.1.2 Le titre

La ressource produite par le pipeline *Gallic(orpor)a* a besoin d'un titre par lequel elle peut être exploitée. L'application `alto2tei` que j'ai créée lui attribue le nom du document source dont le texte la ressource représente. Le schème TEI permet de construire un nouveau titre lors de l'encodage ou d'associer plusieurs titres au document TEI. Cependant, nous avons pris la décision d'emprunter le titre du document source tel qu'il se donne par les sources externes des métadonnées. Au lieu d'aller en détail, dont la TEI se permet, le pipeline simplifie sa manière d'intituler l'encodage en utilisant le même nom associé au document qu'il transcrit.

Dans le schème TEI plusieurs titres peuvent être indiqués dans le `<titleStmt>` afin de détailler plusieurs aspects de l'encodage. Par exemple, le projet *The Bodelian First Folio* a encodé les premières éditions des drames de Shakespeare en TEI et chacun porte plusieurs types de titre. [1] L'encodage de la comédie *Twelfth Night* possède un titre du type "*statement*" qui représente le titre tel qu'il se trouve sur l'imprimé historique (ligne 2, Fig. 2). Il donne aussi une variante du titre qui se trouve également dans la source (ligne 2, Fig. 2). Enfin, l'encodage présente le titre qui sert à identifier la source aux archives, « Bodelian First Folio, Arch. G c.7 » (ligne 4, Fig. 2)

```
1 <titleStmt>
2   <title type="statement">Twelfe Night, or What You Will from Mr.
      William Shakespeares comedies, histories, &amp;amp; tragedies.
      Published according to the true originall copies.</title>
```

1. *The Bodelian First Folio : Digital Facsimile of the First Folio of Shakespeare's Plays.* URL : `http://firstfolio.bodleian.ox.ac.uk/` (visité le 27/08/2022).

```
3   <title type="variant">Mr. VVilliam Shakespeares comedies, histories,
      &amp;amp; tragedies</title>
4   <title type="distinctive">Bodleian First Folio, Arch. G c.7</title>
5 <!-- ... -->
6 </titleStmt>
```

FIGURE 2 – Les informations sur le titre d'un imprimé [2]

Pour un manuscrit, l'attribution d'un titre pourrait obliger la création d'un titre qui ne se trouve pas sur le document source. Par exemple, le projet *CatCor* qui a encodé des lettres écrites par Catherine II de la Russie a choisi d'attribuer aux encodages un titre qui s'appuie sur l'identifiant du document source. Dans l'encodage d'une lettre destinée à Fréderick II le 21 juillet 1744, le `<title>` dans le `<titleStmt>` est un titre qui n'apparaît nulle parte sur la source (ligne 2, Fig. 3). [3] Contrairement à l'encodage de l'imprimé de Shakespeare, l'encodage de la lettre manuscrite ne donne pas de type au titre. La classification d'un élément `<title>` avec l'attribut `@type` n'est pas nécessaire mais elle est recommandée s'il y a plusieurs titres.

```
1 <titleStmt>
2   <title>CatCor Project: letter-02633</title>
3 <!-- ... -->
4 </titleStmt>
```

FIGURE 3 – Les informations sur le titre d'un manuscrit [4]

Afin d'encoder les éléments `<title>` à l'échelle, il faut qu'un logicielle (1) ait d'accès aux métadonnées et (2) sache la nature des titres associés au document source. Certains corpus auront d'accès aux métadonnées déjà classifiées. Le catalogue général de la BnF, par exemple, organise ses métadonnées dans une structure de données XML UNIMARC. Chaque titre associé au document est balisé dans des éléments XML qui portent sur le type du titre. Par exemple, l'UNIMARC présente le type « titre uniforme » dans l'élément `<mxc:datafield tag="500">` et le type « titre de forme » dans l'élément `<mxc:datafield tag="503">`. En récupérant les métadonnées depuis une source ainsi organisée, un logiciel pourrait attribuer un type à l'élément `<title>`. Cependant, ce qui est le cas pour le corpus du projet *Gallic(orpora)*, si certains documents traités n'ont pas de métadonnées accessibles dans un tel format, un logiciel ne peut pas parvenir au tel détail dans le `<titleStmt>`.

1.1.3 L'auteur

Après le titre, le `<titleStmt>` renseigne sur l'individu ou les individus aux lesquels la propriété intellectuelle du document source est attribuée. Cette don-

2. *The Bodelian First Folio*
3. CatCor PROJECT. *Letter 02633 : To Frederick II (the Great), 21 July 1744*. Sous la dir. d'Andrew KAHN et RUBIN-DETLEV. 2021. URL : `https://catcor.seh.ox.ac.uk/id/letter-02633`.
4. ibid.

née est encodée dans l'élément `<author>`. L'encodage peut être si minimal que l'élément `<author>` ne contient que du texte ou de l'élément simple `<name>` qui ensuite contiendrait du texte non annoté. (cf. Figure 4) Sinon d'autres éléments TEI peuvent baliser et apporter plus de détail sur les composants du nom de l'auteur. (cf. Figure 5)

```
<author>
  <name>Donatien Alphonse François de Sade</name>
</author>
```

FIGURE 4 – L'auteur simple

```
<author xmlid="Sa1">
  <persName>
    <forename>Donatien Alphonse François</forename>
    <nameLink>de</nameLink>
    <surname>Sade</surname>
    <ptr type="isni" target="0000000084961458"/>
  </persName>
</author>
```

FIGURE 5 – L'auteur enrichi

Un élément qui descend souvent du `<author>` est le `<ptr>` qui veut dire *pointer* en anglais. Il indique une ressource ou une donnée externe, tel que l'identifiant ISNI, afin d'enrichir les informations de l'objet auquel il est attaché. On voit un exemple du *pointer* sur les lignes 9, 14, et 19 de la Figure 1 où l'élément indique l'ORCID unique de l'individu responsable pour la création de la ressource numérique.

La Figure 5 montre l'exemple d'un document réel TEI produit par le pipeline *Gallic(orpor)a*. Le nom de l'auteur est divisé en trois composants, selon les données UNIMARC fournies par le catalogue général de la BnF. Le catalogue désigne *Donatien Alphonse François de* comme la « partie du nom autre que l'élément d'entrée ». L'UNIMARC balise cette partie secondaire dans l'élément `<mxc:subfield code="b">` de l'élément `<mxc:datafield tag="700">` ou `<mxc:datafield tag="701">`.[5] Mon application `alto2tei`, que j'ai développée pour le pipeline *Gallic(orpor)a*, a ensuite tiré la partie *de* du nom puisqu'elle est un lien entre ses composants et peut donc être encodé dans l'élément TEI `<nameLink>`. Le catalogue général a reconnu le nom *Sade* comme le nom d'entrée de l'auteur, que l'UNIMARC balise dans l'élément `<mxc:subfield code="a">`. Le document TEI du pipeline *Gallic(orpor)a* a donc balisé cette donnée dans l'élément `<surname>`, c'est-à-dire le nom de famille.

Un défi de l'application `alto2tei` était la présentation des données portant sur les auteurs, surtout quand il y en avait plusieurs. Dans le schème TEI, le

5. *Manuel UNIMARC : format bibliographique*. Transition bibliographique - Programme national. URL : `https://www.transition-bibliographique.fr/unimarc/manuel-unimarc-format-bibliographique/` (visité le 27/08/2022).

`<titleStmt>` peut renseigner sur plusieurs auteurs en répétant l'élément comme dans la Figure 6. Le schème TEI permet d'indiquer le rôle de chaque auteur listé dans un `<titleStmt>`. Mais, défaut de métadonnées, l'application `alto2tei` et le pipeline *Gallic(orpor)a* ne s'appuient pas sur cette donnée. Même s'il serait intéressant, toute autrice nommée et tout auteur nommé dans le `<teiHeader>` n'a pas de relation détaillée au document source.

Il y a deux raisons pour ainsi modéliser les données sur les auteurs. Dans un premier temps, l'application `alto2tei` recherche la catégorie d'auteur dans les données renvoyées par l'API IIIF dans ce qui s'appelle le *manifest*. Renvoyées en format JSON, les données du *manifest* sont souvent très minimales, n'ayant que le nom de l'autrice ou de l'auteur. De temps en temps les données du *manifest* IIIF ont des dates de l'individu. Mais le seul aspect sur lequel on peut compter est le nom. Après l'API IIIF, l'application recherche les données dans le catalogue général de la BnF en utilisant son API SRU. Les données du catalogue sont bien plus détaillées, mais le schème UNIMARC ne permet pas d'enregistrer de l'égalité entre plusieurs auteurs ni préciser la nature de leur contribution au document. Dans l'exemple de la Figure 6, les données UNIMARC de la BnF n'indiqueraient pas que Giacomo Meyerbeer est le compositeur de l'opéra *Les Huguenots* ni que lui et le librettiste Eugène Scribe partagent en parts égales la responsabilité pour l'œuvre.

```
<titleStmt>
  <title>Les Huguenots</title>
  <author xml:id="Me1">
    <persName>
      <forename>Giacomo</forename>
      <surname>Meyerbeer</surname>
      <ptr type="isni" target="0000000122817116"/>
    </persName>
  </author>
  <author xml:id="Sc1">
    <persName>
      <forename>Eugène</forename>
      <surname>Scribe</surname>
      <ptr type="isni" target="000000012122970X"/>
    </persName>
  </author>
  <author xml:id="De1">
    <persName>
      <forename>Émile</forename>
      <surname>Deschamps</surname>
      <ptr type="isni" target="0000000122807567"/>
    </persName>
  </author>
  <author xml:id="Ro1">
    <persName>
      <forename>Gaetano</forename>
      <surname>Rossi</surname>
      <ptr type="isni" target="0000000121219499"/>
    </persName>
  </author>
<!-- ... -->
```

```
</titleStmt>
```

FIGURE 6 – Plusieurs auteurs dans un `<titleStmt>`

1.2 La taille de la ressource numérique (`<extent>`)

Lors de la reconnaissance du texte, les modèles HTR génèrent un certain nombre de fichier ALTO. Grâce à l'unité indiqué comme « images » dans l'élément `<measure>`, la taille est calculée facilement à partir du compte de fichiers ALTO produits et ensuite traités pour que leurs données soient contenues dans l'élément `<sourceDoc>` de la ressource numérique.

```
<extent>
  <measure unit="images" n="20"/>
</extent>
```

FIGURE 7 – La taille de la ressource

1.3 La distribution de la ressource numérique (`<publicationStmt>`)

L'élément `<publicationStmt>` contient des données importantes qui renseignent sur la distribution et les droits d'utilisation de la ressource numérique. Toute métadonnée contenue dedans porte sur le contexte de la création de la ressource, aucune sur le document source représenté. Pour toute ressource produite par le pipeline, les données du `<publicationStmt>` ne doivent pas changer si le projet qui a démarré le pipeline. L'exception est la date de la création de la ressource, que l'application `alto2tei` génère automatiquement.

Comme montre la Figure 8, trois données du `<publicationStmt>` peuvent être personnalisées. L'entité reconnu comme l'éditeur (*publisher*) de la ressource peut être changé selon le projet. Dans l'exemple de la Figure 8, le *publisher* est « Gallic(orpor)a ». L'autorité qui l'a financé et qui est civilement responsable pour la ressource est le DataLab de la BnF, que l'élément `<authority>` indique. Enfin, les droits d'utilisation de la ressource sont indiquées par les éléments `<availability>` et `<licence>`.

```
<publicationStmt>
  <publisher>Gallic(orpor)a</publisher>
  <authority>BnF DATALab</authority>
  <availability status="restricted" n="cc-by">
    <licence target="https://creativecommons.org/licenses/by/4.0/"/>
  </availability>
  <date when="2022-07-29"/>
</publicationStmt>
```

FIGURE 8 – Plusieurs auteurs dans un `<publicationStmt>`

1.4 Le document source (`<sourceDesc>`)

Le dernier aspect de la description bibliographique du `<teiHeader>` porte sur le document source dont le texte est représenté. Cet aspect se balise dans l'élément `<sourceDesc>` qui est le dernier élément du `<fileDesc>` que l'application `alto2tei` génère. Cet élément compte sur le plus grand nombre de sources externes afin d'informer une diversité des métadonnées. La description bibliographique de la source porte sur sa création et sur sa conservation actuelle.

1.4.1 La citation bibliographique (`<bibl>`)

Dans un premier temps, la description de la source s'appuie sur les aspects de sa création. En répétant certaines informations du `<titleStmt>`, l'élément `<bibl>` présente (1) les individus auxquels est attribuée la propriété intellectuelle du document, (2) le titre du document source, (3) son lieu de publication, (4) l'éditeur, et (5) la date de publication. L'arborescence du `<bibl>` est montrée dans la Figure 9.

Puisque la notice du catalogue pour le document source montré dans l'exemple de la Figure 9 peut être trouvée par l'application `alto2tei`, les données du `<bibl>` sont plus complète qu'elle auraient été si l'application avait besoin de compter uniquement sur le *manifest* IIIF. La date de publication ou d'apparition, par exemple, représentée dans l'élément `<date>`, porte des attributs bien détaillés. Les données UNIMARC du catalogue général de la BnF indiquent la certitude qu'a la bibliothèque quant à la date déclarée. Au lieu d'extraire uniquement la date des données UNIMARC, l'application `alto2tei` traite ces autres données et détermine la certitude de la date, « low », « medium », ou « high ».

L'exemple de la Figure 9 est d'un imprimé du XVIIIe siècle. Il a donc été publié par un éditeur. L'éditeur, par contre, est indiqué dans les données UNIMARC de la BnF comme « s.n. » qui indique que la bibliothèque n'est pas certain de l'éditeur. L'application `alto2tei` ne peut pas traiter les données dont elle ne se dispose pas ni veut elle transformer la donnée « s.n. » dans un autre format au cas où elle corrompt la signification. Par conséquent, l'élément `<publisher>` contient la donnée du catalogue même si elle indique une manque d'information.

Le `<bibl>` d'un manuscrit contient les mêmes éléments que celui d'un imprimé. Même si un manuscrit n'a pas d'éditeur ni n'est pas publié de la même manière qu'un imprimé, sa citation bibliographique porte les mêmes éléments `<pubPlace>` et `<publisher>`. Normalement, le catalogue général de la BnF n'indiquera pas de donnée pour ces aspects. Quand la donnée n'est pas disponible, l'application `alto2tei` garde toujours l'arborescence généralisée de la ressource numérique, mais elle indique dans l'élément que la donnée n'était pas trouvée.

```
<sourceDesc>
  <bibl>
    <ptr target="http://catalogue.bnf.fr/ark:/12148/cb30369299r"/>
    <author ref="#Re1">
      <persName>
```

```
        <forename>Jean-François</forename>
        <surname>Regnard</surname>
        <ptr type="isni" target="000000012118509X"/>
      </persName>
    </author>
    <author ref="#Du2">
      <persName>
        <forename>Charles</forename>
        <surname>Du Fresny</surname>
        <ptr type="isni" target="0000000140935001"/>
      </persName>
    </author>
    <title>Scènes françoises de la comédie italienne intitulée "la
    Foire S.-Germain" , comme elles ont paru dans les premières
    représentations</title>
    <pubPlace key="FR">Grenoble</pubPlace>
    <publisher>[s.n.]</publisher>
    <date when="1696" cert="high" resp="BNF">1696</date>
  </bibl>
  <msDesc>
<!-- ... -->
  </msDesc>
</sourceDesc>
```

FIGURE 9 – La citation bibliographique (`<bibl>`)

1.4.2 La description de la source physique (`<msDesc>`)

Dernièrement, le document source physique et son fac-similé numérique sont tous les deux indiqués dans l'élément `<msDesc>`. Tandis que la citation bibliographique (`<bibl>`) porte sur le document en tant qu'une œuvre littéraire créée, la *description du manuscrit* (`<msDesc>`) porte sur le document en tant qu'un objet réel dans le monde. Il a deux enfants principaux qui sont montrés dans la Figure 10. Premièrement, l'élément `<msIdentifier>` sert à identifier le document physique ou le fac-similé numérique dans un catalogue quelque part. Deuxièmement, l'élément `<physDesc>` sert à décrire le type du document, soit manuscrit soit imprimé.

L'identification du document traité et encodé est très importante. Par exemple, un imprimé aura plusieurs exemplaires ; chacun pourrait avoir des différences et produirait une transcription distincte. Afin de bien indiquer quel objet de texte a été traité par le pipeline *Gallic(orpor)a*, il faut identifier le document source physique dont le fac-similé numérique les modèles HTR ont saisi. Les éléments `idno` dans le `<msDesc>` indiquent l'identifiant d'un document. L'identifiant du document source physique est l'élément principal descendant du `<msIdentifier>`. L'identifiant du fac-similé numérique de la base de données Gallica est indiqué dans le `<altIdentifier>` puisque le fac-similé est le document dérivé du document source et donc n'est pas le document indiqué par le `<idno>` principal. Le `<idno>` alternatif est l'ARK du fac-similé numérique sur Gallica. Le `<idno>` principal du `<msDesc>` est le cote du document physique selon le catalogue de la Bibliothèque nationale de France.

```
<sourceDesc>
  <bibl>
<!-- ... -->
  </bibl>
  <msDesc>
    <msIdentifier>
      <country key="FR"/>
      <settlement>Paris</settlement>
      <repository>Bibliothèque nationale de France</repository>
      <idno>YF-5877</idno>
      <altIdentifier>
        <idno type="ark">bpt6k1281160s</idno>
      </altIdentifier>
    </msIdentifier>
    <physDesc>
      <objectDesc>
        <p>Texte imprimé</p>
      </objectDesc>
    </physDesc>
  </msDesc>
</sourceDesc>
```

FIGURE 10 – La description de la source (`<msDesc>`)

2 La description non bibliographique (`<profileDesc>`)

Après la description du fichier (`<fileDesc>`) qui porte sur les détails bibliographiques de la ressource numérique, les métadonnées du `<teiHeader>` renseignent sur une description non bibliographique de la ressource. Normalement la description non bibliographique s'informe des langues utilisées dans le texte représenté. Elle peut aussi s'informer des lieux ou personnages référencés dans le texte. Mais ces détails exigent une analyse linguistique, sinon littéraire, du texte. Un jour l'analyse linguistique du pipeline, en particulier la reconnaissance des entités nommées, sera si perfectionnée et si fiable que le `<profileDesc>` de la ressource numérique contiendra les listes de lieux et de noms dans le texte. Actuellement, le `<profileDesc>` renseigne simplement sur la langue du texte, et l'application `alto2tei` que j'ai créée ne peut porter que sur une seule langue. Il est normal que les ingénieurs qui reprennent ce travail après mon stage aillent évoluer notre modélisation actuelle du `<profileDesc>` pour qu'elle ait des données portant sur plusieurs langues du texte ainsi que sur les noms propres et les lieux nommés dans le texte.

La Figure 12 montre l'exemple d'un `<profileDesc>` complété par l'application `alto2tei` dans le cadre du projet *Gallic(orpor)a*. En prenant le même document dont les autres parties de l'encodage se voient dans les Figures 9 et 10, la langue identifiée est français. Cette langue a été récupérée du *manifest* IIIF du fac-similé numérique sur Gallica. Mais quand l'application `alto2tei` peut aussi accéder aux données du catalogue général de la BnF, l'identifiant de la langue est souvent disponible dans les données UNIMARC. L'identifiant « fre » indique le français moderne. Même si la langue identifiée dans le *manifest* IIIF

était encore français, les données UNIMARC du catalogue de la BnF pourrait préciser qu'il est du moyen français, qui est représenté par l'identifiant « frm ».

```
<profileDesc>
  <langUsage>
    <language ident="fre">français</language>
  </langUsage>
</profileDesc>
```

FIGURE 11 – La description non bibliographique de la source (`<profileDesc>`)

3 La description technique (`<encodingDesc>`)

Le dernier composant du `<teiHeader>` est une description technique de l'encodage. Balisée dans l'élément `<encodingDesc>`, la description technique informe de la manière par laquelle l'encodage a été produit. Tout projet scientifique devrait pouvoir reproduire ses résultats. La ressource doit donc attester à la manière par laquelle les résultats de ses prédictions HTR a été faits. L'élément TEI `<appInfo>` renseigne sur le logiciel HTR qui prédit du segment et du texte sur les images numériques. De l'information sur les images et leur origine sont décrites dans les autres éléments du `<teiHeader>`. L'information sur l'appareil qui les a traité est contenu dans le `<appInfo>`.

En plus du logiciel, le `<encodingDesc>` informe du syntaxe par lequel le texte a été encodé en utilisant l'élément `<classDecl>`. La *déclaration des classes* (`<classDecl>`) porte sur les classes visées à structurer et organiser le texte prédit. Dans le cadre du projet *Gallic(orpor)a*, les classes du vocabulaire *SegmOnto* sont attribuées aux lignes de texte et aux zones dans lesquels le texte s'est trouvé. Puisque les lignes de texte et les zones portent les noms du vocabulaire *SegmOnto*, le `<classDecl>` explique fournit une description de chaque classe. La Figure 12 montre comment la description de classe est balisé dans l'élément `<catDesc>` qui lui-même se trouve balisé dans la catégorie (`<category>`) de la classe, soit la ligne soit la zone.

```
<encodingDesc>
  <appInfo>
    <application ident="Kraken" version="3.0.13">
      <label>Kraken</label>
      <ptr target="https://github.com/mittagessen/kraken"/>
    </application>
  </appInfo>
  <classDecl>
    <taxonomy xml:id="SegmOnto">
      <bibl>
        <title>SegmOnto</title>
        <ptr target="https://github.com/segmonto"/>
      </bibl>
      <category xml:id="SegmOntoZones">
        <catDesc xml:id="MainZone">
          <title>MainZone</title>
```

```
          <ptr target="https://segmonto.github.io/gd/gdZ/MainZone"/>
        </catDesc>
        <catDesc xml:id="TitlePageZone">
          <title>TitlePageZone</title>
          <ptr target="https://segmonto.github.io/gd/gdZ/TitlePageZone"
    />
        </catDesc>
<!-- more SegmOnto Zones -->
      </category>
      <category xml:id="SegmOntoLines">
        <catDesc xml:id="DefaultLine">
          <title>DefaultLine</title>
          <ptr target="https://segmonto.github.io/gd/gdL/DefaultLine"/>
        </catDesc>
        <catDesc xml:id="HeadingLine">
          <title>HeadingLine</title>
          <ptr target="https://segmonto.github.io/gd/gdL/HeadingLine"/>
        </catDesc>
<!-- more SegmOnto Lines -->
      </category>
    </taxonomy>
  </classDecl>
</encodingDesc>
```

FIGURE 12 – La description non bibliographique de la source (`<profileDesc>`)

4 La source des métadonneés

Les trois objets de texte conceptuels sont chacun distincts et leurs métadonnées se trouvent depuis les sources différentes. Le pipeline du projet *Gallic(orpor)a* avait donc besoin de diversifier sa manière de récupérer des métadonnées. L'application `alto2tei` a profité de quatre sources externes de données pour renseigner sur les trois documents conceptuels concernés.

1. les métadonnées sur la ressource lexicographique numérique
   — source : fichier de configuration personnalisable
   — format : YAML (*Yet Another Markup Language*)
   — portée : informations administratives portant sur la création de la ressource, y compris les droits d'utilisation et les autorités qui s'en chargent
2. les métadonnées sur le fac-similé numérique
   — source : manifest IIIF renvoyé de l'IIIF API
   — format : JSON (*JavaScript Object Notation*)
   — portée : données bibliographiques qui servent à identifier les exemplaires numériques traités par des modèles HTR
3. les métadonnées sur le document source physique
   — source : réponse à la requête envoyée à l'API SRU (*Search/Retrieve via URL*) de la BnF, qui interroge le catalogue général de la BnF
   — format : UNIMARC XML

— portée : données bibliographiques, y compris la responsabilité du document source, sa création, et sa conservation actuelle

Les métadonnées sur le document source physique s'appuient encore sur une autre source de données afin de récupérer où se trouve l'institution qui héberge le document. Dans le cadre du projet *Gallic(orpor)a*, la réponse à cette question est toujours Paris, puisque la BnF se situe au capital. Mais, afin d'éviter l'encodage dur, le pipeline recherche cette donnée dans la base de données du Système Universitaire de Documentation (SUDOC) puisque la localisation de l'institution hôte du document physique n'est pas encodée dans les données UNIMARC selon l'usage de la BnF.

4.1 Les sources des données du `<titleStmt>`

Dans le `<titleStmt>`, les données du *manifest* IIIF servent à informer de

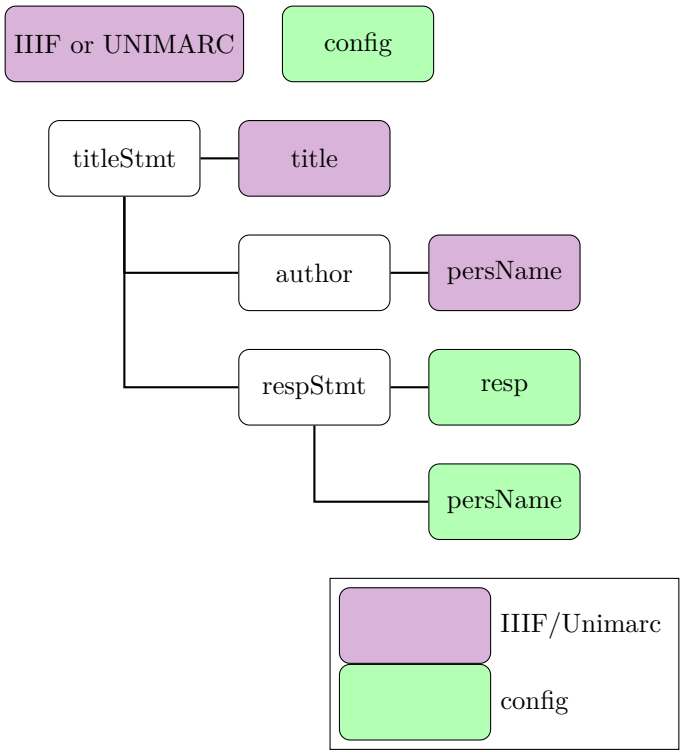

Figure 13

4.2 Les sources des données du `<extent>`

Figure 14

4.3 Les sources des données du `<publicationStmt>`

Figure 15

4.4 Les sources des données du `<sourceDesc>`

Figure 16

4.5 Les sources des données du `<profileDesc>`

Figure 17

4.6 Les sources des données du `<encodingDesc>`

Figure 18