J'ai créé l'application `alto2tei` afin de compléter le pipeline du projet *Gallic(orpor)a*. Pour résumer, la première étape du pipeline est la récupération des pages numérisées du document source. Ensuite il prédit le texte et la mise en page du fac-similé numérique en traitant chaque page avec les modèles HTR. Les données produites par les modèles sont en format XML ALTO. Mais le pipeline veut présenter les données enrichies par les métadonnées et par l'analyse linguistique en format XML TEI puisque le schème TEI est plus utilisé et il convient mieux à l'édition du texte que le schème ALTO. Mon application `alto2tei` a complété le pipeline en construisant un document TEI à partir des données des modèles HTR.

Cependant, les données du text transcrit ne constituent pas tout ce qu'il faut pour construire un document TEI. Il faut aussi des métadonnées qui s'encadrent dans l'élément TEI `<teiHeader>`. L'application `alto2tei` a donc besoin de récupérer les métadonnées à partir des sources externes, qui contiennent des données autre que celles créées par le pipeline. Pour y parvenir, l'application s'appuie sur la technologie d'API (Application Programming Interface) qui lui permet de poser des questions aux sources en ligne et de comprendre leur réponse.

Quelles métadonnées faut-il chercher ? Dans le chapitre **??**, je parle de trois objets de texte conceptuels pertinents : (1) la ressource numérique elle-même en format XML TEI, (2) le fac-similé numérique disponible depuis l'API de la BnF dont les images sont traités par les modèles HTR, et (3) le document physique qui est conservé par la BnF et qui a été numérisé. Chacun est distinct et ses métadonnées se trouvent depuis des sources différentes. Afin d'informer des métadonnées de la ressource numérique, il faut donc s'appuyer sur les métadonnées de ses trois objets de texte.

Le pipeline du projet *Gallic(orpor)a* a besoin de diversifier sa manière de récupérer des métadonnées. L'application `alto2tei` profite de quatre sources externes de données afin de 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, l'application `alto2tei` 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.

1 La récupération des métadonnées

Comme s'explique dans le chapitre **??**, l'application `alto2tei` s'appuie sur un système de fichier fixé. Cela lui permet d'accéder à l'identifiant ARK (Archival Resource Key) du document dont les pages numérisées ont été traitées par les modèles HTR. Cet identifiant donne la clef aux données sur le fac-similé numérique encodées dans le *manifest* IIIF (International Image Interoperability Framework). L'une des données dans le *manifest* IIIF, au moins pour les documents de la base de données Gallica, porte sur la relation entre le fac-similé numérique est la source physique à partir duquel le fac-similé numérique a été produit. Cette relation permet de rechercher les métadonnées du document physique dans le catalogue général de la BnF.

Si la relation entre le document numérique sur Gallica et le document physique dans l'un des magasin des la BnF n'est pas établie, l'application `alto2tei` compte uniquement sur les données du *manifest* IIIF. Ces données sont vérifiées puisqu'il portent sur le fac-similé numérique dont les images les modèles HTR du pipeline ont traité. Cependant, l'application `alto2tei` l'abandonne pas tout essai de produire un document TEI dès que la relation entre le fac-similé numérique et le document physique n'est pas établie. L'application `alto2tei` peut créer deux genres de `<teiHeader>`, l'un avec un maximum d'information et l'autre, au cas où les données du catalogue général de la BnF n'est pas accessible, avec moins d'information. En tout cas, même si la relation entre le fac-similé numérique et le document physique est établie, il peut arriver que les données du catalogue général ne sont pas toutes disponibles et l'application `alto2tei` aura besoin de laisser certains éléments du `<teiHeader>` avec la remarque par défaut, telle que « *Information not available* ».

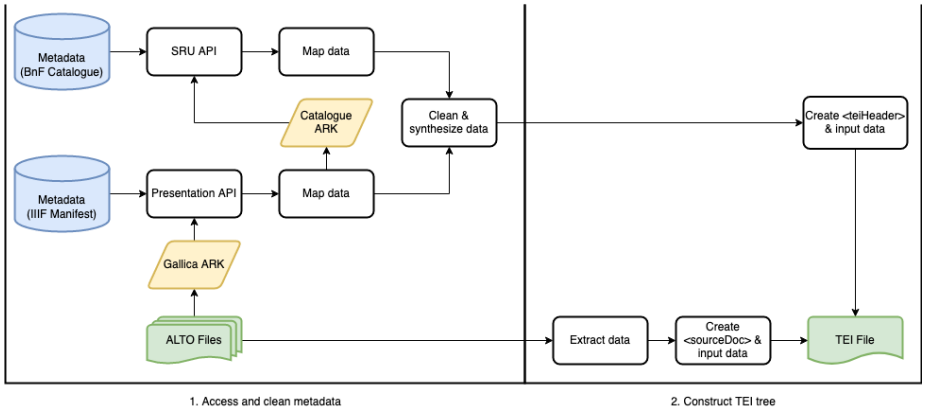

FIGURE 1 – Workflow

2 Les sources des métadonnées

2.1 Les sources des données du `<titleStmt>`

La Figure 2 visualise la source des données du `<titleStmt>`. Le titre est informé soit par le *manifest* IIIF soit par les données UNIMARC de la BnF. D'habitude, l'application `alto2tei` met en priorité les données du catalogue de la BnF. Mais le titre renseigné dans le *manifest* IIIF peut aussi servir à informer du `<title>` du `<titleStmt>` dans le cas où la notice du document dans le catalogue général de la BnF n'était pas trouvée. Les noms des auteurs sont aussi récupérés depuis soit le *manifest* IIIF soit les données du catalogue. Cependant, l'identifiant de l'auteur ainsi que la décomposition du nom en `<forename>` et `<surname>` ne sont trouvés que dans le catalogue général. Si l'application `alto2tei` avait besoin de compter uniquement sur le *manifest* IIIF–à cause de l'inaccessibilité de la notice du document dans le catalogue–les éléments du `<author>` ne seraient pas si détaillés. Les éléments descendants du `<respStmt>` s'appuient sur le fichier de configuration, pour qu'ils puissent être personnalisés par les individus qui utilisent l'application `alto2tei` ou le pipeline *Gallic(orpor)a*.

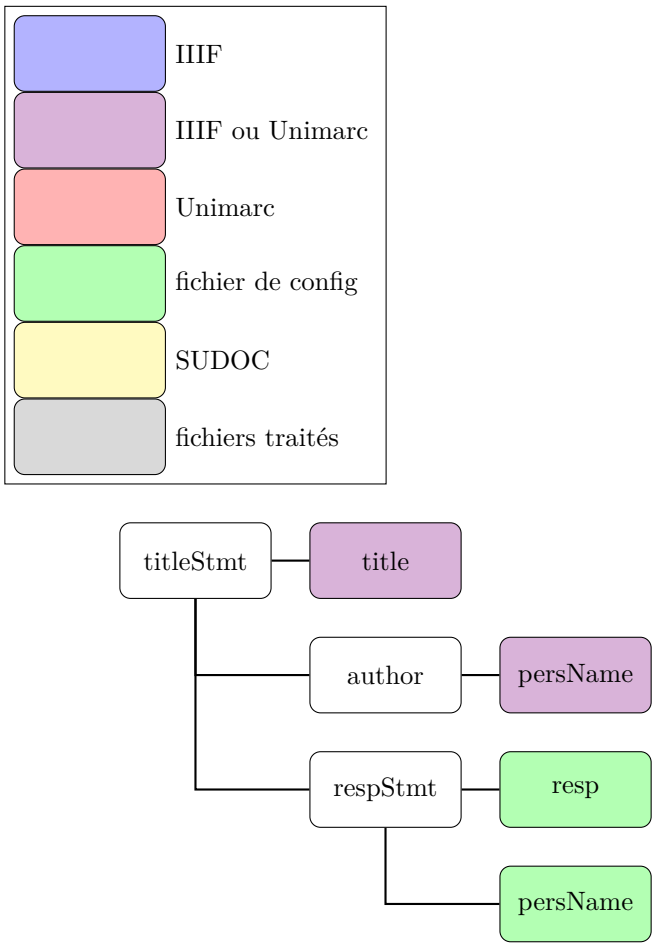

FIGURE 2

2.2 Les sources des données du `<extent>`

L'élément `<extent>` du `<teiHeader>`, qui indique le nombre de fichiers traités par l'application `alto2tei` lors de la création de la ressource numérique, s'appuie simplement sur un compte des fichiers. L'application a besoin de traiter tout fichier ALTO dans le dossier du document, expliqué dans la section **??**, et mettre ses données dans le `<sourceDoc>`, selon la modélisation de l'application `alto2tei`. Elle peut également compter le nombre des fichiers ALTO qui appartiennent au document. Ce compte elle met comme la « taille » de la ressource numérique, dans l'élément `<measure>`.

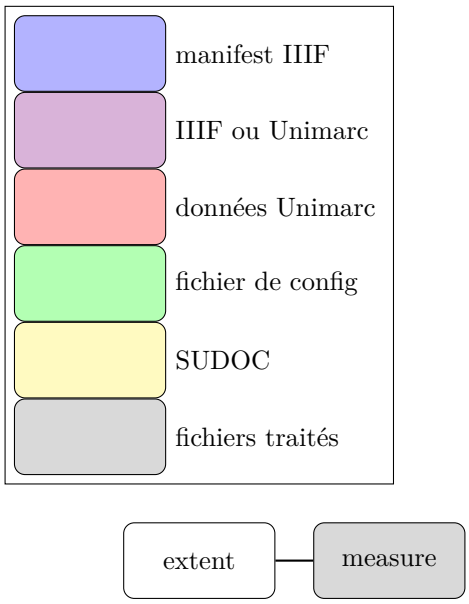

FIGURE 3

2.3 Les sources des données du `<publicationStmt>`

La distribution de la ressource numérique est décrite par l'élément `<publicationStmt>`. Cet élément peut compter uniquement–et idéalement–sur les données du catalogue général de la BnF en format UNIMARC. Mais si l'application a besoin, elle peut remplir certains éléments du `<publicationStmt>` en laissant vide des autres puisque le *manifest* IIIF peut renseigner sur l'auteur du document, son titre, et sa date de création.

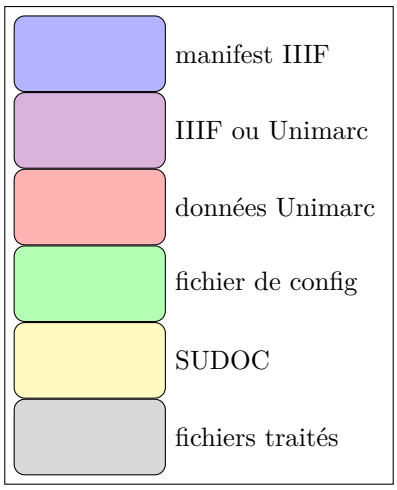

FIGURE 4

2.4 Les sources des données du `<sourceDesc>`

FIGURE 5

2.5 Les sources des données du `<profileDesc>`

FIGURE 6

2.6 Les sources des données du `<encodingDesc>`

FIGURE 7

Parler de la récupération des données depuis les deux APIs discutés (cf. fig. 2.6) ainsi que le fichier YAML de configuration.

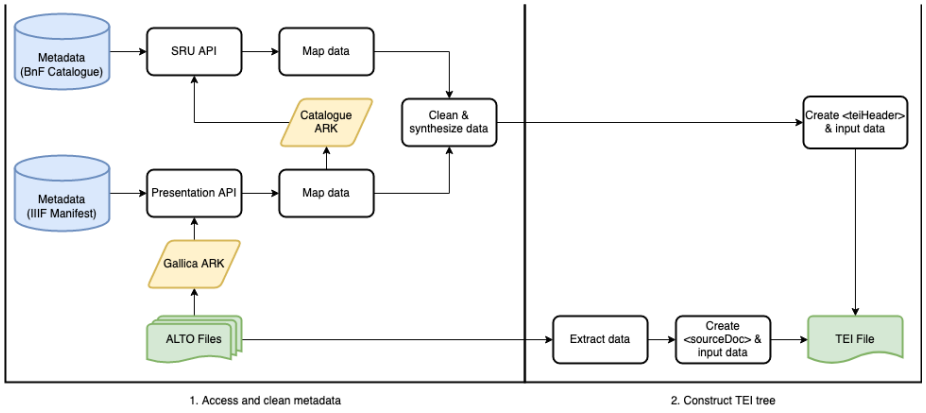

FIGURE 8 – Workflow

2.7 Du manifest IIIF à un dictionnaire Python

Montrer le mapping des données du manifest au dictionnaire Python.

2.8 De l'Unimarc à un dictionnaire Python

Montrer le mapping des données du catalogue au dictionnaire Python.

3 L'analyse des données

Parler de la stratégie d'atténuation des risques en sélectionnant les données fiables. Les métadonnées du catalogue général de la BnF sont utilisées uniquement si le même exemplaire physique du document numérisé sur Gallica a bien été trouvé. Sinon, on risque de mettre les données d'un autre exemplaire de l'oeuvre que celui qui a été transcrit et donc introduire des fausses données, tel que le cote ou même l'éditeur et la date de publication de l'exemplaire.

4 Le modèle du `<teiHeader>`

Montrer le mapping des données au `<teiHeader>`.