# DMP du projet "SourcEncyMe"

*Plan de gestion de données créé à l'aide de DMP OPIDoR, basé sur le modèle "ANR - Modèle de PGD (français)" fourni par Agence nationale de la recherche (ANR).*

## Renseignements sur le plan

| | |
|---|---|
| **Titre du plan** | DMP du projet "SourcEncyMe" |
| **Version** | Version initiale |
| **Domaines de recherche (selon classification de l'OCDE)** | Languages and literature, History and archaeology |
| **Langue** | fra |
| **Date de création** | 2020-10-27 |
| **Date de dernière modification** | 2023-02-27 |
| **Identifiant** | SourcEncyMe PGD oct.2020 |
| **Type d'identifiant** | Identifiant local |
| **Licence** | Creative Commons Attribution Non Commercial No Derivatives 4.0 International |
| **Documents (publications, rapports, brevets, plan expérimental….), sites web associés** | • Site web SourcEncyMe : http://sourcencyme.irht.cnrs.fr/ |

## Renseignements sur le projet

| | |
|---|---|
| **Titre du projet** | SourcEncyMe |
| **Résumé** | SourcEncyMe (*SOURCes des ENCYclopédies MEdiévales*) se veut un outil de référence pour connaître la bibliothèque savante des encyclopédistes, et observer les techniques de compilation médiévales, au moment où l'effort d'assimilation des connaissances antiques et arabes fut le plus important dans l'histoire occidentale médiévale. SourcEncyMe élabore un corpus des encyclopédies médiévales latines, en particulier lors de leur « âge d'or » (1180-1280). L'objectif est d'identifier progressivement les sources textuelles de la vie intellectuelle, puisées par les encyclopédistes dans la tradition antérieure depuis l'Antiquité.<br>Le programme s'attache donc à l'histoire de la transmission des textes grecs, arabes et latins et de la pensée scientifique, théologique et littéraire véhiculée dans les compilations encyclopédiques latines. L'objectif est de mettre en ligne et d'étudier le patrimoine de connaissances regroupé au Moyen Âge sous les noms de « philosophie », « théologie » et « histoire ». L'accent est mis en particulier sur la philosophie naturelle, c'est-à-dire sur la science de la nature, aussi appelée à l'époque « physique ». |

Les encyclopédies médiévales se veulent, de leur propre aveu, une « bibliothèque des bibliothèques », constituée par accumulation de strates successives d'informations tirées de toute la documentation disponible, c'est-à-dire qu'elles offrent un réservoir d'autorités (*auctoritates*). Pour une utilisation aisée par le lecteur médiéval, ces informations se présentent soit sous forme de citations, parfois abrégées, classées par chapitres thématiques, soit dans des catalogues alphabétiques (en particulier pour les *naturalia* comme les plantes ou les animaux). Ces citations sont généralement référencées par un « marqueur de source » que les savants médiévaux, c'est-à-dire le nom de l'autorité, qui se présente souvent comme les noms de l'auteur et de l'œuvre cités : ex : *Aristoteles in libro metheororum*.
C'est la raison pour laquelle SourcEncyMe a adopté une structure fondée sur ce découpage en « unités de citations » précédées par un « marqueur de source ». La structure du site de consultation est donc organisée en arborescence, depuis le découpage en livres et en chapitres, puis en subdivisions adoptées par le compilateur médiéval, jusqu'à « l'unité de base » que constitue la citation précédée de son « marqueur ». Cette dernière peut elle-même être divisée en « segments de citation » au cours de son identification par le chercheur.

Le projet SourcEncyMe, commencé en 2008 et toujours enrichi et en développement, combine les éléments suivants : 1. un corpus annoté de textes encyclopédiques latins balisés en XML-TEI (env. 6 millions de mots) ; 2. un ensemble de méta-données critiques : a. mémentos, i.e. fiches bio-bibliographiques des œuvres et auteurs-sources (*auctoritates*) cités implicitement ou nommément ; b. des identifications des segments de citations ; c. des annotations sur les intermédiaires de transmission (versions du texte, traductions, sources intermédiaires du compilateur, etc.) ; d. une interface de balisage (à l'usage des collaborateurs), permettant d'introduire les noms standardisés des autorités, de compléter les mémentos, d'intégrer les identifications de sources, d'annoter par des commentaires sur la tradition textuelle. Cette interface vient remplacer la précédente interface en ligne, qui permettait aux collaborateurs, via le site, de modifier les données et métadonnées (cet interface avait vieilli et générait des erreurs dommageables).

**Sources de financement**

- Agence nationale de la recherche (ANR) :

**Date de début** 2008-03-01

**Produits de recherche :**

1. Corpus des textes SourcEncyMe (Jeu de données)
2. Identifications des sources et annotations sur la transmission textuelle (Jeu de données)
3. Corpus des mementos auteurs et oeuvres (Jeu de données)
4. Données logicielles : ancienne plateforme collaborative (interface d'administration) et structure de BDD (Logiciel)
5. Données logicielles : environnement de balisage en XML-TEI, XMLmind XML Editor (Logiciel)
6. Données logicielles : site web, scripts de requêtes (Logiciel)

**Contributeurs**

| Nom | Affiliation | Rôles |
|---|---|---|
| Isabelle Draelants - https://orcid.org/0000-0002-2094-9964 | Institut de recherche et d'histoire des textes - CNRS | • Coordinateur du projet<br>• Personne contact pour les données (ex-Plateforme & BDD, Textes, Auteurs & Oeuvres, Identif. & annot.)<br>• Responsable du plan de gestion de données |
| Kuhry Emmanuelle | Institut de recherche et d'histoire des textes | • Personne contact pour les données (Envir. balisage) |
| Seng Henri | Institut de recherche et d'histoire des textes - CNRS | • Personne contact pour les données (Site & scripts) |

# DMP du projet "SourcEncyMe"

## 1. Description des données et collecte ou réutilisation de données existantes

### Corpus des textes SourcEncyMe

**1a. Comment de nouvelles données seront-elles recueillies ou produites et/ou comment des données préexistantes seront-elles réutilisées ?**

Le texte est d'abord OCRisé ou transcrit sous Microsoft Word ou LibreOffice. Il reçoit un pré-balisage à l'aide de styles de caractère et styles de paragraphes qui distinguent :

- la structure et la division en livres, chapitres, titres, unités de citation (styles de paragraphes - Titre 1, Titre 2 etc.)
- les marqueurs médiévaux, c'est-à-dire la référence médiévale à la source (style de caractères)
- éventuellement les sources internes (style de caractères)
- éventuellement les noms propres (style de caractères)

Le fichier XML sous-jacent au fichier de traitement de texte est ensuite récupéré, puis transformé à l'aide d'une feuille de transformation XSLT pour générer un fichier au format XML-TEI.
Les métadonnées suivantes doivent être insérées :

- numérotation des divisions et des unités de citations
- attribution d'un identifiant unique aux unités de citations
- ajout de la balise bibliographique *bibl* en tête de l'unité de citation, contenant le(s) nom(s) canonique(s) (= normalisation des noms d'oeuvres et d'auteurs de la référence médiévale) ainsi que les liens vers les mémentos bio-bibliographiques correspondants

Ceci est réalisé de plusieurs façons possibles et non exclusives :

- via une plateforme collaborative (jusque 2020)
- via un éditeur XML en code brut
- via un éditeur XML avec environnement de balisage WYSIWYG (sous XMLMind XML Editor)

Le document produit est ainsi stocké et interrogé :

- jusque 2020 dans une BDD en PhP-MySQL (contenant des fragments de code XML-TEI)
- depuis 2020, dans une BDD XML native (BaseX)

---

**1b. Quelles données (types, formats et volumes par ex.) seront collectées ou produites ?**

Les textes encyclopédiques (latins) enregistrés dans le corpus SourcEncyMe proviennent, selon les cas, soit de transcriptions de manuscrits médiévaux, soit d'éditions critiques, soit d'éditions imprimées anciennes non critiques.
Les textes sont enregistrés dans un format XML-TEI.
La structure de base est l'unité de citation, délimitée généralement par l'espace qui sépare un marqueur de citation médiéval d'un autre. A l'intérieur de la structure générale soutenue par des éléments div (qui disposent de @n pour la numérotation), on retrouve donc :

- des titres contenus dans *head*
- des unités de citations contenues dans *cit* qui dispose de @n pour la numérotation et @xml:id dont la valeur est un identifiant unique

*cit* se divise en :

- une balise bibliographique *bibl* contenant *ref* (pour les noms d'oeuvres médiévales et modernes) et *author* pour les auteurs
  - *ref* contient le nom canonique et dispose d'un attribut @type="oeuvre" et @target qui pointe vers l'@xml:id du mémento correspondant
  - *author* contient le nom canonique et dispose d'un attribut @ref qui pointe vers l'@xml:id du mémento correspondant
- un élément *quote* qui accueille le texte de la ou des segments de citation

## Identifications des sources et annotations sur la transmission textuelle

**1a. Comment de nouvelles données seront-elles recueillies ou produites et/ou comment des données préexistantes seront-elles réutilisées ?**

Les identifications de sources et annotations scientifiques sont saisies :

- dans un éditeur XML, à la main (fastidieux, risque d'erreur)
- dans la plateforme collaborative (jusque 2020)
- préférentiellement, au sein d'un environnement de balisage en XML-TEI

Ces métadonnées sont saisies par les chercheurs qui en ont la responsabilité intellectuelle, ou par un ingénieur qui s'occupe de la saisie en XML-TEI proprement dite.

---

**1b. Quelles données (types, formats et volumes par ex.) seront collectées ou produites ?**

Les identifications et annotations sont saisies en mode "stand-off embarqué", en format XML-TEI, dans un élément *note* doté de l'attribut *@type,* ayant pour valeur "identification" ou "annotation" :

- identifications : des éléments vides *milestone* encadrent le segment du texte qui est identifié. L'attribut *@target* sur la *note* d'identification pointe vers l'identifiant unique (@xml:id) du premier *milestone*. Le second *milestone* pointe vers ce même @xml:id au moyen de l'attribut *@prev* qui est aussi un pointeur. La note contient :
  - un élément *bibl* contenant *ref* (pour les noms d'oeuvre) et *author* pour les auteurs :
    - *ref* contient le nom canonique et dispose d'un attribut @type="oeuvre" et @target qui pointe vers l'@xml:id du mémento correspondant
    - *author* contient le nom canonique et dispose d'un attribut @ref qui pointe vers l'@xml:id du mémento correspondant
  - un élément p qui contient le texte de l'identification
- annotations : un élément vide *anchor* matérialise la position de l'annotation dans le texte. L'attribut @target sur la *note* d'identification pointe vers l'identifiant unique (@xml:id) de l'élément *anchor*.

## Corpus des mementos auteurs et oeuvres

**1a. Comment de nouvelles données seront-elles recueillies ou produites et/ou comment des données préexistantes seront-elles réutilisées ?**

Un mémento, c'est-à-dire une fiche bio-bibliographique, correspond à chaque auteur et chaque oeuvre citée (environ 1750 fiches pour le moment, voir le produit de recherche "Auteurs et oeuvres").

Elle est destinée à contenir des informations fondamentales et critiques sur la source d'après l'état disponible de l'érudition (appellations d'après des répertoires d'autorités antiques et médiévales et les divers usages ; datation ; contexte), sur son attribution (oeuvre authentique ou pseudépigraphique), des références bibliographiques à l'édition de référence utilisée, des liens vers des référentiels d'autorités et des liens vers des reproductions en ligne (manuscrits médiévaux).

---

**1b. Quelles données (types, formats et volumes par ex.) seront collectées ou produites ?**

Les mémentos sont créés en XML-TEI et stockés dans une base de données XML (BaseX).

Ces fiches de type bio-bibliographique suivent globalement la structure du schéma du pôle "Document numérique" de la MRSH de Caen, auquel ont été ajoutées pour chaque oeuvre ou auteur antique ou médiéval des parties essentielles à l'objectif du projet : oeuvres authentiques, oeuvres pseudépigraphiques, pseudépigraphes, manuscrits de référence, versions de l'oeuvre. Ces parties sont contenues chacune dans un élément *note* de @type="commentaire" et avec un @subtype adapté.

## Données logicielles : ancienne plateforme collaborative (interface d'administration) et structure de BDD

**1a. Comment de nouvelles données seront-elles recueillies ou produites et/ou comment des données préexistantes seront-elles réutilisées ?**

Les données d'identification des sources et d'annotations d'érudition ont été enregistrées via une plateforme collaborative d'administration mise à la disposition des collaborateurs à partir de 2010. Depuis 2020, ces données (informations scientifiques d'érudition) ont été récupérées. Des scripts de migration ont été développé afin de transformer les données de la base de données relationnelles (MySQL) en documents XML et les importer dans une base de données XML (BaseX).

Les nouvelles données seront introduites par une interface de balisage.

---

**1b. Quelles données (types, formats et volumes par ex.) seront collectées ou produites ?**

Une partie des données étaient stockés dans une base de données relationnelles et a été converti en XML. L'autre l'était déjà et a simplement était importé dans BaseX.

Environ 2000 documents XML pour les auteurs et oeuvres.
13 documents XML pour les textes.

## Données logicielles : environnement de balisage en XML-TEI, XMLmind XML Editor

**1a. Comment de nouvelles données seront-elles recueillies ou produites et/ou comment des données préexistantes seront-elles réutilisées ?**

L'environnement de balisage "SourcEncyMe" pour les encyclopédies médiévales est créé à l'aide de l'éditeur XML XMLMind XML Editor sur le modèle de l'environnement de balisage des compilations "Ichtya" développé par le PDN de Caen (https://www.unicaen.fr/recherche/mrsh/document_numerique/outils/compilations).

---

**1b. Quelles données (types, formats et volumes par ex.) seront collectées ou produites ?**

L'environnement de balisage "SourcEncyMe" s'utilise au sein de l'éditeur XML XMLMind XML Editor. Il est constitué de :

- feuilles CSS pour l'affichage des documents selon différentes vues
- fichiers de configuration décrivant les commandes disponibles, qui utilisent le langage Java, mais aussi XSLT et XPath.

Il comprend en outre un plugin *Pluco* pour l'indexation, développé avec Java, par l'équipe du pôle "Document numérique" de la MRSH de Caen (P.-Y. Buard) et configuré pour l'utilisation au sein de l'IRHT par le Pôle numérique de l'IRHT (Henri Seng).

## Données logicielles : site web, scripts de requêtes

**1a. Comment de nouvelles données seront-elles recueillies ou produites et/ou comment des données préexistantes seront-elles réutilisées ?**

Le site est développé en PHP avec le framework Symfony et déployé sur un serveur à l'IRHT.
Il communique avec une base de données XML native BaseX et l'affichage des données XML se fait avec des feuilles de style XSLT. L'interface de consultation utilise des feuilles XSLT pour l'affichage des données XML.
La recherche fonctionne avec Elasticsearch comme moteur de recherche. Les données seront indexées de manière régulière dans ce serveur afin de bénéficier d'une recherche en quasi temps réel.

---

**1b. Quelles données (types, formats et volumes par ex.) seront collectées ou produites ?**

Un site web avec accès au corpus via un masque d'interrogation.

## 2. Documentation et qualité des données

### Corpus des textes SourcEncyMe

**2a. Quelles métadonnées et quelle documentation (par exemple méthodologie de collecte et mode d'organisation des données) accompagneront les données ?**

Un mode d'emploi du site web SourcEncyMe.irht.cnrs.fr et des rubriques d'aide sont disponibles pour les utilisateurs.
Un mode d'emploi de la plateforme collaborative (de 2010 à 2020) à destination des collaborateurs, et depuis 2020 une méthodologie d'encodage fondée sur les principes de la TEI, donnent accès au schéma régissant la structure des textes et des métadonnées.
Dans le document TEI, un en-tête (élément *teiHeader*) permet d'indiquer un certain nombre d'informations utiles :

- informations sur l'édition
- responsabilités intellectuelles
- historique des révisions
- manuscrits ou éditions sources

---

**2b. Quelles mesures de contrôle de la qualité des données seront mises en œuvre ?**

Les textes transcrits ou OCrisés font l'objet d'une relecture avant mise en ligne, qui ne permet pas d'éliminer toutes les erreurs. Le corpus SourcEncyMe est dynamique et évolutif. Des corrections sont apportées au cours de l'analyse et du traitement appliqués à chaque oeuvre latine par les collaborateurs.
Le texte latin est présenté sans son apparat critique et sans l'éventuel commentaire qui l'accompagnait dans l'édition imprimée dont il peut provenir. Il est découpé en citations.
Quand il s'agit d'un texte provenant d'une édition critique, il n'est pas modifié par rapport aux choix philologiques opérés par l'auteur de l'édition critique; des aménagements de forme peuvent y être apportés, qui sont notifiés dans un avertissement (p. ex. standardisation orthographique légère - comme 'j' changé en 'i', 'w' changé en 'uu' -, ou modification de types de parenthèses ou de crochets pour des raisons de balisage). Des corrections éventuelles peuvent être apportées; elles sont alors notifiées et balisées comme telles dans le texte enregistré.
Les textes tirés d'éditions anciennes ont souvent une ponctuation abondante répondant à des normes qui ne sont plus d'actualité ; pour certains d'entre eux, la ponctuation a été allégée.
IMPORTANT : La mise en ligne des textes a pour objectif l'identification des sources et non la mise à disposition critique d'un texte vérifié philologiquement. Les textes ne peuvent donc pas être cités comme tel à partir du corpus : la mise à disposition des textes ne dispense pas l'utilisateur du recours à l'édition de base.

## Identifications des sources et annotations sur la transmission textuelle

**2a. Quelles métadonnées et quelle documentation (par exemple méthodologie de collecte et mode d'organisation des données) accompagneront les données ?**

Une méthodologie d'encodage permet de donner accès au schéma régissant la structure des textes et des métadonnées.

On entend par "sources des textes encyclopédiques" les auteurs et les oeuvres citées par les encyclopédies.

Les identifications donnent accès aux sources "réelles" des textes encyclopédiques, du moins telles qu'elles peuvent être identifiées en fonction de la littérature scientifique disponible aujourd'hui. Des critères d'identification selon la stratigraphie des sources sont mis en place et expliqués aux collaborateurs : identification prioritaire de la source indiquée par le marqueur médiéval, tandis que les annotations sont réservées aux sources intermédiaires.

Un mémento, c'est-à-dire une fiche bio-bibliographique, correspond à chaque auteur et chaque oeuvre citée (environ 1750 fiches pour le moment, voir le produit de recherche "Auteurs et oeuvres"). Elle contient des informations fondamentales et critiques sur : 1. la source (appellations d'après des répertoires d'autorités antiques et médiévales et les divers usages ; datation ; contexte) ; 2. son attribution (oeuvre authentique ou pseudépigraphique), et 3. des références bibliographiques à l'édition de référence utilisée.

Ces sources sont identifiées sous forme de référence à 1. la structure de l'oeuvre citée (livre, chapitre, subdivision, numéro éventuel de citation) et à 2. l'édition de référence, dont la référence est donnée dans le mémento relié à l'oeuvre citée.

---

**2b. Quelles mesures de contrôle de la qualité des données seront mises en œuvre ?**

L'identification des sources et la pose d'annotations scientifiques est sous la responsabilité intellectuelle du chercheur qui en est l'auteur et dépend de son degré de compétence et d'expérience. Ces identifications et annotations peuvent être revues ou corrigées par les collaborateurs successifs et en fonction de l'évolution de l'érudition sur le sujet.

## Corpus des mementos auteurs et oeuvres

**2a. Quelles métadonnées et quelle documentation (par exemple méthodologie de collecte et mode d'organisation des données) accompagneront les données ?**

Dans le document TEI, un en-tête (élément *teiHeader*) permet d'indiquer un certain nombre d'informations utiles :

- créateur de la notice
- historique des révisions

Une méthodologie d'encodage donne accès au schéma régissant la structure des textes et des métadonnées (schéma qui s'inspire, pour les mémentos, du schéma du pôle "Document numérique" de la MRSH de Caen pour l'indexation).

(Voir aussi la rubrique "Textes" 2a).

---

**2b. Quelles mesures de contrôle de la qualité des données seront mises en œuvre ?**

La rédaction des mémentos est sous la responsabilité intellectuelle du chercheur qui en est l'auteur et dépend de son degré de compétence et d'expérience. Ces indications peuvent être revues ou corrigées par les collaborateurs successifs et en fonction de l'évolution de l'érudition sur le sujet.

## Données logicielles : ancienne plateforme collaborative (interface d'administration) et structure de BDD

**2a. Quelles métadonnées et quelle documentation (par exemple méthodologie de collecte et mode d'organisation des données) accompagneront les données ?**

**2b. Quelles mesures de contrôle de la qualité des données seront mises en œuvre ?**

### Données logicielles : environnement de balisage en XML-TEI, XMLmind XML Editor

**2a. Quelles métadonnées et quelle documentation (par exemple méthodologie de collecte et mode d'organisation des données) accompagneront les données ?**

Une méthodologie d'encodage permet de donner accès au schéma régissant la structure des textes et des métadonnées et de décrire le détail des commandes de l'environnement. A l'intérieur des fichiers de configuration, des commentaires permettent d'expliquer les parties du code.
Cette méthodologie est mise à la disposition des collaborateurs.

**2b. Quelles mesures de contrôle de la qualité des données seront mises en œuvre ?**

La "propreté" du code pourra être soumise à évaluation auprès d'informaticiens.

### Données logicielles : site web, scripts de requêtes

**2a. Quelles métadonnées et quelle documentation (par exemple méthodologie de collecte et mode d'organisation des données) accompagneront les données ?**

Un mode d'emploi pour les utilisateurs est disponible sur le site et explique ses différentes fonctions.
Les métadonnées sont au standard TEI.

**2b. Quelles mesures de contrôle de la qualité des données seront mises en œuvre ?**

Les informations scientifiques introduites par les collaborateurs (spécialistes et compétents dans un domaine de la transmission des textes) sont vérifiées et susceptibles d'être mises à jour par les collaborateurs qui reviendront sur l'information (p. ex. pour les identifications de sources ou la rédaction de fiches-mémentos bio-bibliographiques).

## 3. Stockage et sauvegarde pendant le processus de recherche

### Corpus des textes SourcEncyMe

**3a. Comment les données et les métadonnées seront-elles stockées et sauvegardées tout au long du processus de recherche ?**

Les textes et métadonnées sont stockées dans une BDD XML BaseX sur un serveur de l'Institut de recherche et d'histoire des textes (IRHT) au Pôle numérique d'Orléans.

Chaque chercheur(e) souhaitant éditer ou modifier un texte ou ajouter des métadonnées, récupère une copie du fichier grâce à un système de *versioning (Gitlab)* - le fichier une fois édité est renvoyé sur le serveur central pour mettre à jour la version existante par le moyen de scripts.

---

**3b. Comment la sécurité des données et la protection des données sensibles seront-elles assurées tout au long du processus de recherche ?**

L'accès aux données "versionnées" nécessite un compte utilisateur au gitlab de l'IRHT.

## Identifications des sources et annotations sur la transmission textuelle

**3a. Comment les données et les métadonnées seront-elles stockées et sauvegardées tout au long du processus de recherche ?**

Les textes et métadonnées sont stockées dans une BDD XML BaseX sur un serveur de l'IRHT ; chaque chercheur(e) souhaitant éditer un texte récupère une copie du fichier grâce à un système de *versioning (Gitlab)* - le fichier une fois édité est renvoyé sur le serveur central pour mettre à jour la version existante.

---

**3b. Comment la sécurité des données et la protection des données sensibles seront-elles assurées tout au long du processus de recherche ?**

L'accès aux données versionnées nécessite un compte utilisateur au gitlab de l'IRHT.

## Corpus des mementos auteurs et oeuvres

**3a. Comment les données et les métadonnées seront-elles stockées et sauvegardées tout au long du processus de recherche ?**

Les textes et métadonnées sont stockés dans une BDD XML BaseX sur un serveur de l'IRHT ; chaque chercheur(e) souhaitant éditer un mémento récupère une copie du fichier grâce à un plugin permettant l'indexation au moyen de référentiels d'autorités : le plugin Pluco développé par l'équipe du pôle "Document numérique" de la MRSH Caen.

---

**3b. Comment la sécurité des données et la protection des données sensibles seront-elles assurées tout au long du processus de recherche ?**

L'édition des métadonnées nécessitera la configuration du plugin Pluco avec des identifiants.

## Données logicielles : ancienne plateforme collaborative (interface d'administration) et structure de BDD

**3a. Comment les données et les métadonnées seront-elles stockées et sauvegardées tout au long du processus de recherche ?**

---

**3b. Comment la sécurité des données et la protection des données sensibles seront-elles assurées tout au long du processus de recherche ?**

## Données logicielles : environnement de balisage en XML-TEI, XMLmind XML Editor

**3a. Comment les données et les métadonnées seront-elles stockées et sauvegardées tout au long du processus de recherche ?**

L'environnement sera prochainement mis à disposition sur un serveur de l'IRHT.

---

**3b. Comment la sécurité des données et la protection des données sensibles seront-elles assurées tout au long du processus de recherche ?**

Le projet ne contient pas de données sensibles.

## Données logicielles : site web, scripts de requêtes

**3a. Comment les données et les métadonnées seront-elles stockées et sauvegardées tout au long du processus de recherche ?**

Le site est développé avec le framework Symfony et déployé sur un serveur à l'IRHT.
Les données de la base de données XML sont indexées dans un serveur Elasticsearch qui fournit un moteur de recherche et permet d'effectuer des recherches quasiment en temps réel.

---

**3b. Comment la sécurité des données et la protection des données sensibles seront-elles assurées tout au long du processus de recherche ?**

Le site est principalement une interface de consultation seule. Il est envisagé que certains contenus textuels (en dehors des données) puissent être modifiables depuis le site.

# 4. Exigences légales et éthiques, codes de conduite

## Corpus des textes SourcEncyMe

**4a. Si des données à caractère personnel sont traitées, comment le respect des dispositions de la législation sur les données à caractère personnel et sur la sécurité des données sera-t-il assuré ?**

Le projet ne contient pas de données à caractère personnel, à part le nom et prénom des collaborateurs dans les données XML, c'est-à-dire des collaborateurs qui pourraient avoir introduit des données (auteurs de fiches par exemple) .

---

**4b. Comment les autres questions juridiques, comme la titularité ou les droits de propriété intellectuelle sur les données, seront-elles abordées ? Quelle est la législation applicable en la matière ?**

Le texte latin est présenté dans le corpus sans son apparat critique et sans l'éventuel commentaire qui l'accompagnait dans l'édition imprimée dont il peut provenir.
L'identité de l'auteur de la transcription ou de l'édition critique du texte médiéval et la date de la transcription ou de l'édition sont indiquées dans la référence.
Les éditions dont le texte latin est repris et mis en ligne le sont selon les modalités suivantes, selon les cas :

- imprimés anciens libres de droit
- éditions modernes libres de droit
- accord obtenu auprès de l'auteur de l'édition critique et/ou de l'éditeur commercial
- travail de recherche fourni directement par l'auteur

Les collaborateurs (chercheurs) qui apportent des identifications de sources sont identifiés dans chaque enregistrement de métadonnées.

---

**4c. Comment les éventuelles questions éthiques seront-elles prises en compte, les codes déontologiques respectés ?**

Les données sont en accès public et émanent de la recherche publique.
Les codes de conduite nationaux et internationaux s'appliquent, ainsi que la déontologie de la recherche publique et en particulier du CNRS.
Toute utilisation ou citation des données de la recherche produite par SourcEncyMe doit faire l'objet d'une référence par l'utilisateur.
Toute récupération ou retraitement des données doit être soumise au responsable du projet et pourra, le cas échéant, faire l'objet d'un partenariat ou d'une convention de recherche.

## Identifications des sources et annotations sur la transmission textuelle

**4a. Si des données à caractère personnel sont traitées, comment le respect des dispositions de la législation sur les données à caractère personnel et sur la sécurité des données sera-t-il assuré ?**

Le projet ne contient pas de données à caractère personnel, à part le nom et prénom des collaborateurs dans les données XML, c'est-à-dire des collaborateurs qui pourraient avoir introduit des données (auteurs de fiches-mémentos ou d'identifications par exemple).

---

**4b. Comment les autres questions juridiques, comme la titularité ou les droits de propriété intellectuelle sur les données, seront-elles abordées ? Quelle est la législation applicable en la matière ?**

Les identifications et annotations sont signées par leurs auteurs (@resp), comme toute modification ultérieure.

**4c. Comment les éventuelles questions éthiques seront-elles prises en compte, les codes déontologiques respectés ?**

(Voir rubrique "Textes" 4c.)

## Corpus des mementos auteurs et oeuvres

**4a. Si des données à caractère personnel sont traitées, comment le respect des dispositions de la législation sur les données à caractère personnel et sur la sécurité des données sera-t-il assuré ?**

Le projet ne contient pas de données à caractère personnel, à part le nom et prénom des collaborateurs dans les données XML, c'est-à-dire des collaborateurs qui pourraient avoir introduit des données (auteurs de fiches-mémentos ou d'identifications par exemple).

**4b. Comment les autres questions juridiques, comme la titularité ou les droits de propriété intellectuelle sur les données, seront-elles abordées ? Quelle est la législation applicable en la matière ?**

Les mémentos sont signés par leurs auteurs (responsabilités intellectuelles décrites dans le *teiHeader* et @resp sur les notes), comme toute modification ultérieure.
(Voir aussi rubrique "Textes" 4b).

**4c. Comment les éventuelles questions éthiques seront-elles prises en compte, les codes déontologiques respectés ?**

(voir rubrique "Textes" 4c.)

## Données logicielles : ancienne plateforme collaborative (interface d'administration) et structure de BDD

**4a. Si des données à caractère personnel sont traitées, comment le respect des dispositions de la législation sur les données à caractère personnel et sur la sécurité des données sera-t-il assuré ?**

L'ancienne base de données ne contenait comme données à caractère personnel que le nom et prénom des collaborateurs dans les données XML, c'est-à-dire de ceux qui pourraient avoir introduit des données (auteurs de fiches-mémentos ou d'identifications par exemple).

**4b. Comment les autres questions juridiques, comme la titularité ou les droits de propriété intellectuelle sur les données, seront-elles abordées ? Quelle est la législation applicable en la matière ?**

**4c. Comment les éventuelles questions éthiques seront-elles prises en compte, les codes déontologiques respectés ?**

## Données logicielles : environnement de balisage en XML-TEI, XMLmind XML Editor

**4a. Si des données à caractère personnel sont traitées, comment le respect des dispositions de la législation sur les données à caractère personnel et sur la sécurité des données sera-t-il assuré ?**

Il n'y a pas de traitement de données à caractère personnel dans le cadre des outils comme l'environnement de balisage.

---

**4b. Comment les autres questions juridiques, comme la titularité ou les droits de propriété intellectuelle sur les données, seront-elles abordées ? Quelle est la législation applicable en la matière ?**

Les fichiers de configuration mentionnent les responsabilités intellectuelles.

---

**4c. Comment les éventuelles questions éthiques seront-elles prises en compte, les codes déontologiques respectés ?**

Sans objet.

## Données logicielles : site web, scripts de requêtes

**4a. Si des données à caractère personnel sont traitées, comment le respect des dispositions de la législation sur les données à caractère personnel et sur la sécurité des données sera-t-il assuré ?**

Il n'y a pas de traitement de données à caractère personnel dans les scripts.

---

**4b. Comment les autres questions juridiques, comme la titularité ou les droits de propriété intellectuelle sur les données, seront-elles abordées ? Quelle est la législation applicable en la matière ?**

Non pertinent.

---

**4c. Comment les éventuelles questions éthiques seront-elles prises en compte, les codes déontologiques respectés ?**

Le projet ne suscite pas de questions éthiques.
La déontologie en vigueur est la même que pour toute publication scientifique au CNRS et dans la recherche publique.

## 5. Partage des données et conservation à long terme

### Corpus des textes SourcEncyMe

**5a. Comment et quand les données seront-elles partagées ? Y-a-t-il des restrictions au partage des données ou des raisons de définir un embargo ?**

Les données (textes + métadonnées critiques) ont été segmentées, c'est-à-dire qu'elles sont mises en ligne et affichées une citation à la fois et non sous forme de texte séquentiel. Cette option a été adoptée à dessein pour qu'on ne puisse télécharger l'intégralité d'une oeuvre enregistrée.

Toute intention de téléchargement intégral d'une oeuvre pour en retraiter les données est soumise à l'autorisation du responsable du projet ; elle fera l'objet d'une collaboration sous convention.

Les données peuvent être modifiées dans le temps, si un collaborateur juge utile ou nécessaire de les corriger ou de les compléter.

Un colophon et des mentions légales donnent ces indications sur le site SourcEncyMe.

---

**5b. Comment les données à conserver seront-elles sélectionnées et où seront-elles préservées sur le long terme (par ex. un entrepôt de données ou une archive) ?**

Nous explorons les solutions proposées par Huma-Num, comme NAKALA, pour la conservation des données, en partenariat avec le CINES.

---

**5c. Quelles méthodes ou quels outils logiciels seront nécessaires pour accéder et utiliser les données ?**

Un navigateur à jour.

---

**5d. Comment l'attribution d'un identifiant unique et pérenne (comme le DOI) sera-t-elle assurée pour chaque jeu de données ?**

Chaque texte dispose d'un identifiant unique @xml:id sur l'élément englobant *TEI*.

Les identifiants ne sont pas générés, ils sont transformés à la "main" sur la base du passage du nom canonique en minuscules, et des espaces en "_" et autres règles de ce type.

Ex : "Epistole Horatii" en "epistole_horatii" (pour un memento oeuvre)
"Speculum historiale, version SM trifaria (Ms Douai BM 797)" en "speculum_historiale_version_sm_trifaria_ms_douai_bm_797" (pour un texte).

Cette génération pourra être configuré sur les éditeurs XML.

De manière plus générale, si les données sont déposées dans un dépôt de données tel que Nakala (voir “Corpus des textes SourcEncyMe” 5b), un DOI leur sera attribué automatiquement.

### Identifications des sources et annotations sur la transmission textuelle

**5a. Comment et quand les données seront-elles partagées ? Y-a-t-il des restrictions au partage des données ou des raisons de définir un embargo ?**

Les données sont dynamiques et en constante évolution en fonction du temps qui peut être consacré au projet et de l'évolution de la recherche et de l'érudition en général.

Les données (textes + métadonnées critiques) sont mis en ligne et affichés une citation à la fois. A dessein, il n'a pas été prévu de

possibilité de télécharger l'intégralité d'un corpus.
Les données peuvent être modifiées dans le temps, si un collaborateur juge utile ou nécessaire de les corriger ou de les compléter.
(Voir rubrique "Textes" 5a)

**5b. Comment les données à conserver seront-elles sélectionnées et où seront-elles préservées sur le long terme (par ex. un entrepôt de données ou une archive) ?**

Nous explorons des solutions avec Huma-Num pour la conservation des données, en partenariat avec le CINES.

**5c. Quelles méthodes ou quels outils logiciels seront nécessaires pour accéder et utiliser les données ?**

Un navigateur.

**5d. Comment l'attribution d'un identifiant unique et pérenne (comme le DOI) sera-t-elle assurée pour chaque jeu de données ?**

L'@xml:id de chaque identification/annotation est construit par l'env. de balisage en fonction de la position dans l'arborescence du fichier.
De manière plus générale, si les données sont déposées dans un dépôt de données comme cela est envisagé (voir “Corpus des textes SourcEncyMe” 5b), un DOI leur sera attribué automatiquement.

## Corpus des mementos auteurs et oeuvres

**5a. Comment et quand les données seront-elles partagées ? Y-a-t-il des restrictions au partage des données ou des raisons de définir un embargo ?**

Le contenu (données) du projet n'est pas statique, il évolue constamment.
Les données (textes + métadonnées critiques) sont mis en ligne et affichés une citation à la fois. Il n'y a pas de possibilité de télécharger l'intégralité d'un corpus. Les données peuvent être modifiées dans le temps, si un collaborateur juge utile ou nécessaire de les corriger ou de les compléter.
(Voir aussi rubrique "Textes" 5a.)

**5b. Comment les données à conserver seront-elles sélectionnées et où seront-elles préservées sur le long terme (par ex. un entrepôt de données ou une archive) ?**

**5c. Quelles méthodes ou quels outils logiciels seront nécessaires pour accéder et utiliser les données ?**

Un navigateur.

**5d. Comment l'attribution d'un identifiant unique et pérenne (comme le DOI) sera-t-elle assurée pour chaque jeu de données ?**

Chaque texte dispose d'un identifiant unique @xml:id sur l'élément englobant *TEI*.
Les identifiants ne sont pas générés, ils sont transformés à la "main" sur la base du passage du nom en minuscules, et des espaces en "_" et autres règles de ce type.
Ex : "Epistole Horatii" en "epistole_horatii" (pour un memento oeuvre)
"Speculum historiale, version SM trifaria (Ms Douai BM 797)" en "speculum_historiale_version_sm_trifaria_ms_douai_bm_797" (pour un texte).

## Données logicielles : ancienne plateforme collaborative (interface d'administration) et structure de BDD

**5a. Comment et quand les données seront-elles partagées ? Y-a-t-il des restrictions au partage des données ou des raisons de définir un embargo ?**

Non pertinent parce que l'ancienne base de données a été remplacée.

---

**5b. Comment les données à conserver seront-elles sélectionnées et où seront-elles préservées sur le long terme (par ex. un entrepôt de données ou une archive) ?**

Une copie de l'ancienne base de données (non accessible) est conservée en archives sur le serveur de l'IRHT à Orléans.

---

**5c. Quelles méthodes ou quels outils logiciels seront nécessaires pour accéder et utiliser les données ?**

Non pertinent (BDD obsolète en 2021).

---

**5d. Comment l'attribution d'un identifiant unique et pérenne (comme le DOI) sera-t-elle assurée pour chaque jeu de données ?**

Non pertinent, l'ancienne BDD ne sert plus depuis fin 2021.

## Données logicielles : environnement de balisage en XML-TEI, XMLmind XML Editor

**5a. Comment et quand les données seront-elles partagées ? Y-a-t-il des restrictions au partage des données ou des raisons de définir un embargo ?**

L'environnement de balisage pour les encyclopédies médiévales et sa méthodologie seront mis à disposition sur l'un des sites de l'IRHT.

---

**5b. Comment les données à conserver seront-elles sélectionnées et où seront-elles préservées sur le long terme (par ex. un entrepôt de données ou une archive) ?**

L'environnement sera stocké sur les serveurs de l'IRHT.

**5c. Quelles méthodes ou quels outils logiciels seront nécessaires pour accéder et utiliser les données ?**

La version 8 de l'éditeur XML XMLMind XML Editor.

**5d. Comment l'attribution d'un identifiant unique et pérenne (comme le DOI) sera-t-elle assurée pour chaque jeu de données ?**

Sans objet.

### Données logicielles : site web, scripts de requêtes

**5a. Comment et quand les données seront-elles partagées ? Y-a-t-il des restrictions au partage des données ou des raisons de définir un embargo ?**

Le code source du site est stocké et versionné dans le Gitlab de l'IRHT. L'accès au code source est soumis à l'autorisation du responsable de projet.

**5b. Comment les données à conserver seront-elles sélectionnées et où seront-elles préservées sur le long terme (par ex. un entrepôt de données ou une archive) ?**

Pour l'instant, l'archivage du code-source n'est pas prévu sous une autre forme que sur le GitLab de l'IRHT.

**5c. Quelles méthodes ou quels outils logiciels seront nécessaires pour accéder et utiliser les données ?**

L'accès est soumis à l'autorisation du responsable de projet.

**5d. Comment l'attribution d'un identifiant unique et pérenne (comme le DOI) sera-t-elle assurée pour chaque jeu de données ?**

Il n'y a pas d'identifiant unique pour les scripts. Non pertinent.

## 6. Responsabilités et ressources en matière de gestion des données

### Corpus des textes SourcEncyMe

**6a. Qui (par exemple rôle, position et institution de rattachement) sera responsable de la gestion des données (c'est-à-dire le gestionnaire des données) ?**

Responsabilité scientifique : Isabelle Draelants (Directrice de recherche, IRHT-CNRS), porteuse du projet SourcEncyMe.

Responsabilité technique : Pôle numérique de l'IRHT (resp. Cyril Masset, Ingénieur de recherche IRHT-CNRS).

Responsabilité éditoriale : François Bougard, directeur de l'IRHT.

Plan de gestion de données rédigé et mis à jour par I. Draelants, E. Kuhry (historiennes médiévistes) et complété par Cyril Masset et Henri Seng (informaticiens).

Les données sont saisies par divers collaborateurs ou partenaires (chercheurs, ingénieurs, techniciens) impliqués dans les phases successives du projet sous la coordination du porteur principal. Ils sont parfois engagés sous contrats courts en fonction des financements qui ont pu être obtenus.

---

**6b. Quelles seront les ressources (budget et temps alloués) dédiées à la gestion des données permettant de s'assurer que les données seront FAIR (Facile à trouver, Accessible, Interopérable, Réutilisable) ?**

Le coût global est impossible à chiffrer. Il est très considérable en temps de travail depuis le lancement du projet, sans parler des données tirées du travail antérieur produit dans les deux laboratoires hébergeurs successifs (une sorte de capital d'expérience et de travail).

Le coût en terme de temps de travail s'intègre d'une part aux activités des chercheurs en poste. Il est nettement supérieur au temps de travail indiqué en pourcentage pour les divers collaborateurs et contractuels dans les dossiers soumis aux financeurs successifs, puisqu'une grande part du travail n'a pas été financé. Pour ces raisons, donner un total des montants qui ont fait l'objet d'un financement ne serait pas indicatif.

Certains collaborateurs (post-doctorants, ingénieurs, stagiaires) ont été engagés sous contrats courts en fonction des financements obtenus selon les cas d'agences (ANR), d'institutions (MSH à Nancy, MRSH à Caen, Université de Nancy, laboratoire CMJS à Nancy, IRHT à Paris, Cerlam puis Craham à Caen), de régions (région Lorraine), de groupements de laboratoires (Labex HAstec) ou de projets nationaux englobants (Equipex Biblissima).

Pour l'instant, les frais d'entrepôt ou d'archivage ne sont pas identifiés car ils entrent dans le coût environné des laboratoires hébergeurs qui sont des institutions de recherche publique.

## Identifications des sources et annotations sur la transmission textuelle

**6a. Qui (par exemple rôle, position et institution de rattachement) sera responsable de la gestion des données (c'est-à-dire le gestionnaire des données) ?**

Responsabilité scientifique : Isabelle Draelants (DR, IRHT-CNRS)

Responsabilité technique : Pôle numérique de l'IRHT (resp. Cyril Masset, IR IRHT-CNRS)

Responsabilité éditoriale : François Bougard, directeur de l'IRHT

Des collaborations scientifiques ponctuelles sont liées au traitement (identification et transmission) d'une source particulière par un spécialiste, historien ou latiniste, de cette matière (thématique).

(Voir l'onglet "textes" 6a.)

---

**6b. Quelles seront les ressources (budget et temps alloués) dédiées à la gestion des données permettant de s'assurer que les données seront FAIR (Facile à trouver, Accessible, Interopérable, Réutilisable) ?**

(Voir l'onglet "Textes" 6b).

## Corpus des mementos auteurs et oeuvres

**6a. Qui (par exemple rôle, position et institution de rattachement) sera responsable de la gestion des données (c'est-à-dire le gestionnaire des données) ?**

Responsabilité scientifique : Isabelle Draelants (DR, IRHT-CNRS)
Responsabilité technique : Pôle numérique de l'IRHT (resp. Cyril Masset, IR IRHT-CNRS)
Responsabilité éditoriale : François Bougard, directeur de l'IRHT
Rédaction du plan de gestion de données : I. Draelants et E. Kuhry.

---

**6b. Quelles seront les ressources (budget et temps alloués) dédiées à la gestion des données permettant de s'assurer que les données seront FAIR (Facile à trouver, Accessible, Interopérable, Réutilisable) ?**

(voir rubrique "Textes" 6b)

## Données logicielles : ancienne plateforme collaborative (interface d'administration) et structure de BDD

**6a. Qui (par exemple rôle, position et institution de rattachement) sera responsable de la gestion des données (c'est-à-dire le gestionnaire des données) ?**

Responsabilité scientifique : Isabelle Draelants (DR, IRHT-CNRS)
Responsabilité technique : Pôle numérique de l'IRHT (resp. Cyril Masset, IR IRHT-CNRS; développeur : Henri Seng)
Responsabilité éditoriale : François Bougard, directeur de l'IRHT

---

**6b. Quelles seront les ressources (budget et temps alloués) dédiées à la gestion des données permettant de s'assurer que les données seront FAIR (Facile à trouver, Accessible, Interopérable, Réutilisable) ?**

## Données logicielles : environnement de balisage en XML-TEI, XMLmind XML Editor

**6a. Qui (par exemple rôle, position et institution de rattachement) sera responsable de la gestion des données (c'est-à-dire le gestionnaire des données) ?**

Responsabilité scientifique : Isabelle Draelants (DR, IRHT-CNRS)
Responsabilité technique : Pôle numérique de l'IRHT (resp. Cyril Masset, IR IRHT-CNRS)
Responsabilité éditoriale : François Bougard, directeur de l'IRHT
Production de l'environnement de balisage : Emmanuelle Kuhry (Post-doctorante et Ingénieur de recherche IRHT)

---

**6b. Quelles seront les ressources (budget et temps alloués) dédiées à la gestion des données permettant de s'assurer que les données seront FAIR (Facile à trouver, Accessible, Interopérable, Réutilisable) ?**

## Données logicielles : site web, scripts de requêtes

**6a. Qui (par exemple rôle, position et institution de rattachement) sera responsable de la gestion des données (c'est-à-dire le gestionnaire des données) ?**

Responsabilité scientifique : Isabelle Draelants (DR, IRHT-CNRS)
Responsabilité technique : Service informatique du Pôle numérique de l'IRHT (resp. Cyril Masset, IR IRHT-CNRS).
Responsabilité éditoriale : François Bougard, directeur de l'IRHT

---

**6b. Quelles seront les ressources (budget et temps alloués) dédiées à la gestion des données permettant de s'assurer que les données seront FAIR (Facile à trouver, Accessible, Interopérable, Réutilisable) ?**

Pour le moment, aucun moyen financier ou autre n'est disponible pour s'assurer que les données seront FAIR.
Cf. réponse 5a, onglet "Textes" qui explique pourquoi le corpus n'est (volontairement) pas téléchargeable.