Biblissima+

Observatoire des cultures écrites
de l'argile à l'imprimé

Plan de gestion de données de l'infrastructure Biblissima+, observatoire des cultures écrites anciennes de l'argile à l'imprimé (ANR-21-ESRE-0005)

Version 2.0

Octobre 2023

Table des matières

V.II Livrables

V.III Fiches pratiques

I. Introduction

1 Résumé

Ce document décrit le plan de gestion des données (PGD) de l'observatoire des cultures écrites anciennes **Biblissima+**. Fédérant 16 établissements et une entreprise privée, réunis en un consortium, **Biblissima+** crée une infrastructure numérique multipolaire de recherche fondamentale et de service consacrée à l'histoire de la transmission des textes anciens, des premières tablettes d'argile mésopotamiennes aux premiers livres imprimés, sur tous les supports et dans toutes les langues. Biblissima+ concerne donc l'ensemble des collections patrimoniales transmettant des textes anciens, y compris les sources archéologiques, les sceaux et monnaies, mais aussi les archives d'érudits modernes et de chercheurs contemporains quand elles apportent des informations originales sur les textes anciens et leur circulation.

Le présent document correspond à la version 2 du livrable à fournir à l'ANR dans les 24 mois suivant le démarrage du projet. Il s'appuie en partie sur le modèle générique diffusé par l'ANR en 2019. Il prend en compte la grille de relecture des PGD proposée par l'INIST-CNRS[1] en 2020.

Le document s'organise autour de 3 parties principales.

- La première partie définit de manière générale les grands principes directeurs de gestion des données qui s'appliquent à la collecte, le stockage, l'organisation, le partage et l'archivage des données. Ces principes ont volontairement été définis de manière minimale, afin de favoriser l'harmonisation des pratiques de manière transversale à l'ensemble du projet et de rester compatible avec les politiques de données définies par la ou les organisations tutelle(s) des équipes impliquées.
- La seconde partie présente une vue d'ensemble de toutes les données et codes sources qui seront produits. Elle tient aussi compte de l'infrastructure technique mise en place pendant la première période de financement ÉquipEx dont **Biblissima+** prend le relais. Cette vue d'ensemble est fournie sous forme de tableaux synthétiques, faisant ressortir les choix opérés en matière de dépôt dans des entrepôts tiers, afin de permettre la citation, le partage et l'archivage des données.
- La troisième partie s'attache à décrire de manière plus détaillée le périmètre des données et des codes source dont l'équipe technique de **Biblissima+** est en charge (portail Biblissima, ses différents sites web et l'infrastructure technique sous-jacente). Ce périmètre est en effet placé sous la responsabilité directe de l'établissement porteur Campus Condorcet. Les livrables produits par les équipes partenaires sont quant à eux sous la responsabilité de leurs tutelles.

Le PGD sera actualisé en continu pour refléter l'évolution des décisions ou faire l'inventaire des jeux de données produits au fil du temps. Une version stabilisée du PGD sera par la suite fournie à l'ANR tous les deux ans jusqu'en 2029.

Biblissima+ bénéficie d'une aide de l'Etat gérée par l'ANR au titre du Programme d'investissements d'avenir intégré à France 2030, portant la référence ANR-21-ESRE-0005.

1. Diffusée sur la plateforme de ressources d'autoformation DORANUM le 27/07/2020 - DOI : 10.13143/R7GM-6C38. ↵

2 Définition et objectifs d'un plan de gestion des données

Un plan de gestion des données (PGD) permet de consigner dans un document centralisé toutes les informations importantes sur les données d'un projet de recherche. Il décrit la manière dont elles seront traitées au long de leur cycle de vie, de l'étape initiale de collecte ou de production, à l'étape finale de publication ou d'archivage, en passant par les différents stades de gestion proprement dite comme le classement, la structuration, le stockage, le traitement ou analyse à l'aide d'outils numériques… Dans un contexte de généralisation du numérique au sein des activités de recherche, l'enjeu est de limiter les risques d'obsolescence technologique, de perte ou de gaspillage de ressources afin de garantir que les données sous-jacentes aux résultats scientifiques pourront être effectivement diffusées, partagées et réutilisées, que ce soit au bénéfice d'autres recherches ou de la société tout entière. Le but d'un PGD est avant tout de faciliter et d'optimiser cette gestion en permettant l'anticipation des actions de structuration et de description. Les données jouent en effet un rôle clé pour l'intégrité scientifique. Leur conservation sous une forme intelligible et exploitable numériquement requiert une planification d'actions concrètes qui est le but du PGD.

Au cours du projet de recherche, le PGD est actualisé en continu pour enregistrer les informations concrètes liées à la mise en œuvre des décisions. En conséquence, son rôle et ses utilisations évoluent dans le temps. Au début du projet, sa fonction principale est de fournir un outil de pilotage et d'aide à la décision. Il s'agit de guider les choix opérationnels de « curation », autrement dit la combinaison d'actions d'intendance et de documentation leur permettant de conserver leur intelligibilité et utilité, aussi longtemps que nécessaire. À la fin du projet de recherche, le PGD s'est enrichi de toutes les informations concrètes sur ce qui a été fait. Il prend aussi en compte les nécessaires adaptations ou réorientations par rapport à ce qui avait été prévu. Il devient alors un élément essentiel pour comprendre le contexte de production des données et en constitue l'historique documenté, afin de fournir les conditions nécessaires à la vérification de l'intégrité scientifique des résultats ou à la production de nouvelles recherches à partir d'elles.

Un PGD doit expliciter de manière synthétique les choix techniques, juridiques et organisationnels concernant les données. 6 dimensions principales sont à prendre en compte :

- L'identification des principaux produits de recherche créés ou collectés durant le projet, principalement numériques : regroupements ou sets de données dits « jeux de données », codes sources de logiciels et scripts, protocoles, méthodes et procédures, etc. ;
- La définition des modalités d'organisation et de description normalisées via des métadonnées ;
- Le stockage et la sécurité des données ;
- Les objectifs concernant la diffusion, le partage ou l'archivage final, en expliquant le degré d'ouverture choisi, les durées de rétention visées, la nécessité ou non d'un archivage numérique pérenne ;
- Les informations sur les questions éthiques, la présence de données sensibles, la justification des restrictions à la diffusion ouverte (open data) ;

- Les ressources prévues pour la mise en œuvre du plan (moyens financiers ou temps de travail des chercheurs et ingénieurs impliqués).

Un PGD est enfin le fruit d'une coopération inter-métiers entre scientifiques, informaticiens, documentalistes, archivistes et éditeurs qui peuvent participer à des degrés divers à son élaboration et à sa mise à jour. Quand elle est correctement mise en œuvre, cette dimension collaborative du PGD peut jouer un rôle particulièrement utile pour contribuer à forger une vision commune au sein d'une équipe. Le PGD offre en effet, sous l'angle spécifique des données et des procédures numériques, une vision d'ensemble du projet à la fois synthétique et concrète. Il peut ainsi faciliter grandement l'intégration des nouveaux collaborateurs ou collaboratrices. Il peut aussi servir de document de référence pour transmettre de manière efficace des informations clé pour comprendre le socle technique du projet.

3 Champ d'application et politiques d'établissements applicables

Le PGD de **Biblissima+** s'applique à toutes les données et jeux de données produits dans le cadre des activités de recherche décrites au sein des livrables du projet. Il porte sur les activités donnant lieu à des versements financiers aussi bien que sur les opérations inscrites dans le projet au titre des apports des établissements et organismes partenaires[1].

3.1 Définition : données et jeux de données

Le terme « données » tel qu'il est utilisé dans ce document doit être entendu au sens large. Dans un programme d'infrastructure de service et de recherche tel que **Biblissima+**, la notion recouvre aussi bien les textes et des corpus de textes, les collections d'images ou de photos, les modèles numériques 3D, les données d'entraînement en intelligence artificielle que les enregistrements audiovisuels ou les bases de données. Elle recouvre également l'ensemble des codes sources, des méthodes et des protocoles qui seront utilisés pour présenter, analyser, instrumenter ou diffuser ces artefacts.

Un « jeu de données » (*dataset* en anglais) peut être défini comme une collection de fichiers électroniques présentant une certaine « unité » et qui sont rassemblés pour former un tout cohérent. L'échelle à laquelle l'agrégation est réalisée ainsi que les critères utilisés sont laissés à l'appréciation des scientifiques. Ces critères peuvent en effet varier de manière importante selon les questions de recherche, la nature des données, les équipements utilisés, ou encore les réutilisations possibles.

3.2 Politiques de science ouverte applicables

Dans leurs activités liées au projet, les équipes de recherche doivent respecter les exigences de la politique de l'ANR en matière de Science ouverte[2].

Dans le cadre de sa politique de science ouverte, l'ANR demande que les projets qu'elle finance ou opère produisent des données dont les modes de structuration et de diffusion respectent 4 principes fondamentaux génériques rassemblés sous l'acronyme « FAIR », à savoir : Facilement trouvables, Accessibles, Interopérables et Réutilisables. L'agence demande également que leur diffusion soit ouverte ou autrement dit sans entrave, en appliquant le principe « aussi ouvert que possible, aussi fermé que nécessaire ». Ainsi, si la mise à disposition sous licence ouverte n'est pas obligatoire, les restrictions à celle-ci ou les délais (embargos) doivent être expliqués ou justifiés dans le PGD.

Les équipes partenaires doivent également respecter les politiques particulières de leur(s) établissement(s) tutelle(s) en la matière.

Partenaire	Document(s) formalisant une politique de Science ouverte	URL ou DOI
Campus Condorcet	-	-
CNRS	Feuille de route Science ouverte (2019) et Plan données (2020)	https://www.science-ouverte.cnrs.fr/
Université PSL	Charte Science ouverte (2020)	https://www.psl.eu/sites/default/files/Charte_science_ouverte_Universite-psl_Mai_2020.pdf
EHESS	-	-
ENS de Lyon	Feuille de route Science ouverte 2023	https://www.ens-lyon.fr/sites/default/files/2023-10/FdR_SO_2023.pdf
MNHN	-	-
Avignon Université	Charte pour la Science ouverte (2023)	https://univ-avignon.fr/wp-content/uploads/2023/06/Charte-pour-la-science-ouverte-Approuvee-en-CR-13.04.23.pdf
Université de Caen	Schema de gouvernance Science ouverte en Normandie	https://science-ouverte.normandie-univ.fr/science-ouverte/science-ouverte-en-normandie/
Université Lyon 3	Charte pour la Science ouverte (2020)	https://www.univ-lyon3.fr/medias/fichier/charte-science-ouverte-lyon3-web_1608112950703-pdf
Université Lyon 2	Feuille de route pour la Science ouverte (2022)	https://www.univ-lyon2.fr/universite/actualites-universitaires/une-feuille-de-route-pour-la-science-ouverte
Université de Poitiers	Feuille de route pour la Science ouverte (2022)	https://www.univ-poitiers.fr/wp-content/uploads/sites/10/2022/09/Science-ouverte-de-lUP.pdf
Université de Tours	Charte pour la science ouverte (2022)	https://www.univ-tours.fr/medias/fichier/charte-science-ouverte_1662627501636-pdf

Partenaire	Document(s) formalisant une politique de Science ouverte	URL ou DOI
SIAF	-	-
TEKLIA	-	-

3.3 Accord de consortium

L'accord de consortium du programme EquipEx Biblissima+ a été signé le 19 avril 2023 par l'ensemble des établissements et organismes partenaires du projet.

Il comporte un article spécifiquement centré sur la politique de science ouverte du programme. Le paragraphe ci-dessous en reproduit l'intégralité.

Accord de consortium Equipex+ BIBLISSIMA+ p. 19

Article 10 – Science ouverte

Les Parties s'accordent sur le principe que les Résultats ne générant pas de droits de propriété intellectuelle ni un savoir-faire secret peuvent être largement diffusés, dans le respect de l'article 8 relatif à la confidentialité, conformément aux politiques de science ouverte nationale et européenne encourageant une mise à disposition ouverte des données, des codes sources et des méthodes sous-jacents aux résultats à des fins d'intégrité scientifique, de réutilisation et d'encouragement de l'innovation.

Conformément à l'article 9 de la Convention attributive ANR, les Parties s'engagent à déposer les publications scientifiques (texte intégral) issues du Projet dans une archive ouverte, soit directement dans HAL soit par l'intermédiaire d'une archive institutionnelle locale, dans les conditions de l'article 30 de la Loi « Pour une République numérique »

Dès que cela sera possible eu égard aux dispositions relatives aux Informations Confidentielles et à la protection et l'exploitation des Résultats, les Parties s'engagent, en vertu du principe d'ouverture par défaut auxquelles elles adhèrent, à favoriser la diffusion large au public des connaissances, données et codes sources issus du Projet.

Les Parties établiront et tiendront à jourun plan de gestion desdonnées dans lequel elles définiront et justifieront ce qui devra rester confidentiel et pour quelle durée (voir article « Publications »), les conditions d'archivage des données et des informations relatives au Projet, et les informations et données qui pourront être diffusées au public ainsi que les modalités de cette diffusion.

Les Parties s'engagent à le mettre à jour pendant l'exécution et à la fin du Projet, conformément à l'article 8 de la Convention attributive d'aide ANR.

Cette clause ne fait en tout état de cause pas obstacle à la protection des Résultatspar un Droit de propriété intellectuelle et, le cas échéant, par la délivrance d'un titre de propriété industrielle.

---

1. Cf. le document de soumission et ses annexes. ↩
2. Cf. https://anr.fr/fr/lanr/engagements/la-science-ouverte/ et le plan d'action 2022 de l'ANR, version 1.1a du 12 octobre 2021 ↩

4 Spécificités du projet Biblissima+

4.1 Présentation générale

L'observatoire des cultures écrites anciennes **Biblissima+** est un projet d'infrastructure numérique consacrée à l'histoire de la transmission des textes produits de l'Antiquité à la Renaissance en Orient comme en Occident, quel qu'en soit le support et quelle qu'en soit la langue. Il crée un portail national offrant un accès unique et simple à des ressources électroniques hétérogènes (documentation écrite originale, collections d'images numérisées de sources, bibliographie et archives de la recherche la concernant). Il constitue également un environnement de travail proposant des chaînes d'outils pour enrichir, partager, réutiliser les corpus. Le but est de permettre des recherches nouvelles sur l'histoire de la transmission des textes et des bibliothèques reposant sur une méthodologie de traitement des données et des codes sources conformes aux objectifs de Science ouverte.

Biblissima+ fédère 16 établissements et une entreprise privée. Il réunit plusieurs équipes de recherche travaillant sur les textes, de l'Antiquité à l'édition numérique, une entreprise et le ministère de la Culture. Il fait partie des équipements structurants pour la recherche ÉquipEx+ sélectionnés en 2020 dans le cadre des Investissements d'avenir. L'équipe chargée du portail Biblissima+ proprement dit est hébergée par le Campus Condorcet, établissement porteur de l'ÉquipEx+. Les équipes partenaires, qui développent les contenus mis en interopérabilité ou diffusés via le portail (ressources scientifiques et outils innovants) sont organisées autour de 7 domaines d'innovation numérique et d'expertise ou « clusters ». Un système d'appels à projets ouvert à tous est destiné à produire de nouveaux jeux de données interopérables et de nouveaux outils à partir d'opérations conjointes de recherche, de documentation, de numérisation et de valorisation portant sur des collections historiques de manuscrits, d'imprimés anciens ou d'autres objets portant du texte.

Biblissima+ s'appuie sur les réalisations et l'expérience de l'ÉquipEx Biblissima (*Bibliotheca bibliothecarum novissima* : observatoire du patrimoine écrit du Moyen Âge et de la Renaissance, 2012-2021). Il hérite de l'infrastructure informatique mise en place pour gérer le portail Biblissima, de sa plateforme de référentiels *data.biblissima*, son moteur de recherche *IIIF-Collections* et de son service *IIIF360* opéré avec le Campus Condorcet et Huma-Num. Il a pour objectif principal de maintenir et développer cette infrastructure et d'étendre potentiellement ses contenus à toutes les langues anciennes et à leurs supports. Il a aussi pour mission de veiller à leur intégration par les communautés par le partage des outils et des pratiques.

4.2 Organisation

Le projet s'articule autour de deux volets principaux.

Le premier (*volet A*), est centré sur la maintenance et le développement de l'infrastructure portail, de ses moteurs de recherche et de son référentiel, épine dorsale de l'infrastructure et composant clé des opérations de mise en interopérabilité. Un de ses principaux enjeux est la définition et la mise en œuvre de mécanismes génériques et stables d'agrégation et d'enrichissement de ressources. Ces mécanismes doivent être capables d'agréger les nouveaux types de données sans nuire à l'efficacité et à la simplicité d'un portail unique. Ils doivent aussi tenir compte des contraintes liées au besoin de s'articuler avec d'autres grandes infrastructures pour certains types de données, notamment la bibliographie ou s'adapter à des sources de données qui sont issues de bases de données évolutives. Il s'agit en somme de mettre au point un « système de mise à jour » en lien étroit avec les communautés notamment parce que toutes les dimensions ne peuvent être automatisées.

Le *volet B* regroupe toutes les contributions financées par le projet et développées par les équipes partenaires au sein des clusters. Dans ces 7 domaines d'innovation numérique, les communautés de chercheurs, les ingénieurs, conservateurs, étudiants partagent les questions, les outils, les standards et inventent de nouveaux outils. Tous reçoivent des moyens pour leurs recherches et leurs développements, mais aussi pour des rencontres annuelles : les *semaines des clusters*. De plus, les résultats, les questions, les idées des clusters sont mis en commun chaque année lors des *Journées Biblissima+*, qui permettent de faire dialoguer les clusters entre eux et de réfléchir au bon chaînage des outils. Ces journées sont couplées avec le Conseil scientifique international annuel, de façon à favoriser les interactions, l'approfondissement, la naissance d'idées nouvelles.

Les 7 domaines d'expertise de Biblissima+ sont organisés selon le cycle de travail sur les sources :

- **Cluster 1** – Acquisition des corpus de sources interopérables (images 2D et 3D) ;
- **Cluster 2** – Prise en compte et cherchabilité des données d'analyse des matériaux ;
- **Cluster 3** – Intelligence artificielle, reconnaissance de formes et d'écritures manuscrites ;
- **Cluster 4** – Traitement approfondi des systèmes graphiques et analyse des documents ;
- **Cluster 5** – Edition de sources selon les standards EpiDoc (pour l'épigraphie : cluster 5a) et TEI (pour les différentes typologies textuelles : cluster 5b) ;
- **Cluster 6** – Défis du patrimoine musical et MEI ;
- **Cluster 7** – Interopérabilité et analyse des textes.

4.3 Particularités de la gestion des données dans Biblissima+

4.3.1 Biblissima+ : un projet « FAIR by design »

La raison d'être de **Biblissima+** étant d'offrir un portail d'accès unifié mettant en interopérabilité collections patrimoniales, archives de la recherche et littérature scientifique, le projet a appliqué les principes FAIR dès le départ et la première période de financement. La diffusion ouverte des données et métadonnées ainsi que le développement open source des outils numériques reste au cœur du positionnement scientifique et technique de **Biblissima+.** Les résultats de l'ÉquipEx, qu'il s'agisse des données descriptives de collections patrimoniales ou d'éditions, des référentiels d'autorité utilisés pour les décrire, d'outils et protocoles développés pour assurer le fonctionnement de l'infrastructure numérique seront diffusés avec des licences les plus ouvertes possibles (CC BY ou Licence ouverte Etalab 2.0), afin de favoriser l'accroissement de leur réutilisation et de leur rayonnement.

4.3.2 3 périmètres de données et de responsabilités à distinguer

L'organisation du projet permet de distinguer 3 périmètres de données en relation avec le statut des équipes productrices au sein du projet. On distingue ainsi les trois périmètres de données, qui sont aussi des périmètres de responsabilités :

- **Périmètre P1** : l'infrastructure logicielle du portail d'accès unifié et ses briques fonctionnelles ;
- **Périmètre P2** : les contributions des équipes partenaires dans le cadre des livrables du projet, qui constituent les autres « briques » de l'écosystème de ressources et d'outils de **Biblissima+** ;
- **Périmètre P3** : les résultats d'opérations conjointes de recherche, de documentation, de numérisation et de valorisation financées après sélection de l'appel à manifestation d'intérêt[1].

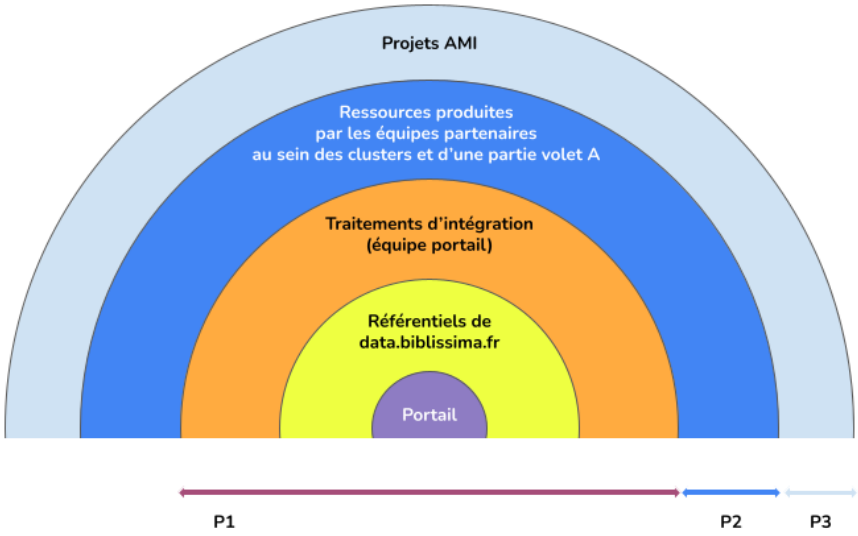

Figure 1: Périmètres de données du projet Biblissima+

4.3.3 PGD principal et PGDs particuliers

Étant donné l'ampleur du projet, la variété et l'hétérogénéité des données qui seront produites, le PGD de **Biblissima+** définit des lignes directrices et des principes de choix qui s'appliquent aux 3 périmètres mais ne détaille que la gestion des données du périmètre géré et développé par l'équipe technique des trois ingénieurs employés par l'établissement porteur Campus Condorcet.

4.3.4 Contenu du PGD pour chaque périmètre de données

Le tableau suivant détaille chacun de ces périmètres et précise sa relation au PGD.

Périmètre	Définition	Contenu	Relation avec le PGD et dépôt du document
P1	Infrastructure numérique (livrables pilotés par l'équipe portail au sein du volet A du projet)	Interfaces web du portail / Moteurs de recherche / Cluster de données / Plateforme de référentiels d'autorité / Protocoles et scripts d'ingestion de données et de mise à jour des agrégations	Contenu du PGD principal, rédigé et mis à jour par l'équipe Portail sous la responsabilité du bureau exécutif. Chaque version obligatoire dans le cadre de la convention avec l'ANR est déposée dans la communauté Zenodo (sous licence libre Etalab 2.0). Une version "vivante" est publiée en ligne dans le système de publication de documentation Mkdocs (voir en ligne : https://dmp.biblissima.fr/).
P2	Autres livrables des volets A et B, Outils de la boîte à outils de Biblissima	Bibliothèques numériques / Catalogues et répertoires / Bases de données scientifiques / Corpus spécialisés / Éditions de textes / Outils de traitement scientifique des corpus / Tutoriels et vidéos de formation	PGDs autonomes par livrables rédigés par leurs responsables scientifiques et techniques ou sous leur responsabilité. Ces documents ont vocation à être déposés dans la communauté Zenodo de Biblissima+, au plus tard à la fin du programme Biblissima+. Il est recommandé de le rendre librement consultable, mais ce n'est pas obligatoire.
P3	Productions liées aux opérations financées dans le cadre d'un appel à projets annuel	Opérations conjointes de recherche, de documentation, de numérisation et de valorisation portant sur des collections historiques de manuscrits, d'imprimés anciens, ou d'autres objets portant du texte, associant au moins un établissement de conservation et un établissement	Un PGD est demandé dans le dossier de soumission de l'AMI. Ce document a vocation à être déposé dans la communauté Zenodo de Biblissima+ à l'échéancé dé la convéntion. Il est recommandé de le rendre librement consultable, mais ce n'est pas obligatoire. Seul l'accès pour les membres de Biblissima+ est requis.

Périmètre	Définition	Contenu	Relation avec le PGD et dépôt du document
		d'enseignement et/ou de recherche.	

1. Voir la page web: https://projet.biblissima.fr/fr/appels-a-projets/projets-partenariaux-biblissima ↩

II. Le PGD de Biblissima+ (cadre de référence)

5 Le plan de gestion des données de Biblissima+

5.1 Objectifs

Le plan de gestion des données de **Biblissima+** a pour objectif d'établir une stratégie globale pour la gestion des données créées et collectées durant le projet (2021-2029). La démarche proposée vise en particulier à faciliter le partage et l'archivage de jeux de données accessibles et réutilisables au sein d'entrepôts de confiance dédiés à ces fonctions.

Le PGD est un livrable officiel du projet d'ÉquipEx.

Le présent document correspond à la version 2 du PGD qui est à fournir à l'ANR dans les 24 mois après le démarrage officiel, conformément à la convention attributive d'aide de l'ANR signée du 26 octobre 2021.

Le présent document aborde les points suivants :

- Les principes directeurs de gestion et de diffusion des données s'appliquant à l'ensemble des périmètres de données identifiés au sein du projet (cf. supra) ;
- Les recommandations minimales à appliquer par chaque équipe partenaire de manière à favoriser l'harmonisation des pratiques ;
- Une vue d'ensemble des politiques de partage et d'archivage pour l'ensemble des futurs livrables, proposée sous forme de tableaux de synthèse (périmètres P2 et P3) ;
- Un PGD détaillé de l'infrastructure numérique du périmètre P1 (utilisant le modèle de PGD de l'ANR) ;
- Des annexes.

6 Lignes directrices

Les lignes directrices du PGD de **Biblissima+** définissent des exigences minimales à respecter dans la gestion des données produites ou collectées durant le projet.

6.1 1

Le PGD de **Biblissima+** est centré sur le périmètre de données lié à l'infrastructure numérique gérée par l'équipe technique du projet appéle également « équipe portail ».

Ce périmètre (P1) comprend le portail web, le cluster de données sous-jacent, les référentiels d'autorité assurant la mise en interopérabilité des ressources ainsi que les outils de traitement mis en œuvre.

2

Les briques de l'infrastructure produites par les équipes partenaires (périmètre P2) ou les opérations financées via les appels à projet annuels (périmètre P3) font l'objet de PGDs particuliers, rédigés et mis à jour par les responsables scientifiques et techniques de ces contributions.

Le livre blanc[1] de **Biblissima+** comprend une grand variété de contributions. À titre d'illustration, on peut mentionner les types de contributions suivants :

- les catalogues de notices,
- les extractions de bases de données scientifiques,
- lescorpus spécialisés,
- les éditions de sources,
- les thésaurus,
- les listes d'autorité (noms de personnes, de lieux, identifiants),
- des vocabulaires contrôlés
- ainsi que les codes sources de logiciels ou scripts ou modèles informatiques (3D, intelligence artificielle, apprentissage machine) associés aux outils, méthodes et protocoles proposés.

6.2 3

Les jeux de données du périmètre P1 sont placés sous la responsabilité du bureau exécutif et de l'établissement porteur Campus Condorcet.

Les jeux de données des périmètres P2 et P3, sont quant à eux sous la responsabilité des équipes partenaires produisant ou collectant les données et les codes sources et de leurs établissements tutelles.

6.3 4

Les PGDs particuliers sont rédigés et mis à jour par les responsables scientifiques des livrables et sont placés sour la responsabilité des établissements tutelles des équipes scientifiques impliquées. Il doivent respecter les politiques de Science ouverte de ces tutelles quand elles existent. Pour rédiger ces documents il est recommandé d'utiliser un modèle de PGD permettant l'export des données dans le format normalisé défini par l'organisation internationale RDA. Deux outils sont disponibles à ce jour : le « modèle structuré » de plan de gestion de données de la plateforme DMP OpiDor[2] (en français) et l'outil en ligne ARGOS, proposé par l'infrastructure européenne OpenAIRE[3] (en anglais)[4].

6.4 5

Le PGD principal aussi bien que les PGDs particuliers s'inscrivent dans une démarche de Science ouverte conforme à la politique générale de l'ANR en la matière et au plan national porté par le MESRI[5]. Les données produites doivent être structurées et rendues exploitables en fonction des principes FAIR (faciles à trouver, accessibles, interopérables, réutilisables).

Le PGD principal et les PGDs particuliers définissent explicitement la manière dont ces données seront préservées et partagées. Ils indiquent a minima : l'entrepôt qui sera utilisé pour le dépôt, le niveau d'agrégation, les conditions d'accès et les licences de réutilisation.

6.5 6

Les données sont mises à disposition de la manière la plus ouverte possible. Lorsqu'il n'est pas possible de diffuser les données sous une licence ouverte, ou lorsque la diffusion ouverte est soumise à un embargo, les raisons en sont expliquées dans le PGD (droits de propriété intellectuelle, présence de données sensibles, etc.)

6.6 7

Tout jeu de données ayant vocation à être intégré dans le portail **Biblissima+** doit faire l'objet d'un dépôt documenté dans un entrepôt fournissant un identifiant pérenne (par exemple un DOI).

Le dépôt contient a minima 3 fichiers au format `.txt` ou `.md`[7] :

- un fichier README[8] expliquant notamment son organisation (arborescence des fichiers) et le dictionnaire des données ;

- un fichier LICENCE donnant la licence de diffusion, précisant les conditions de réutilisation en général et en particulier au sein du portail **Biblissima+** ;
- un fichier AUTHORS contenant la liste des auteurs et des contributeurs éventuels.

Le diagramme ci-dessous illustre les principales étapes du processus et les responsabilités associées. Les étapes liées aux périmètres P2 et P3 sont sous la responsabilité des équipes qui définissent, extraient et organisent leur « jeu de données ». À titre d'exemple on peut citer : une collection de notices issues d'un catalogue, d'une base de données scientifiques, d'un corpus TEI ou d'un référentiel.

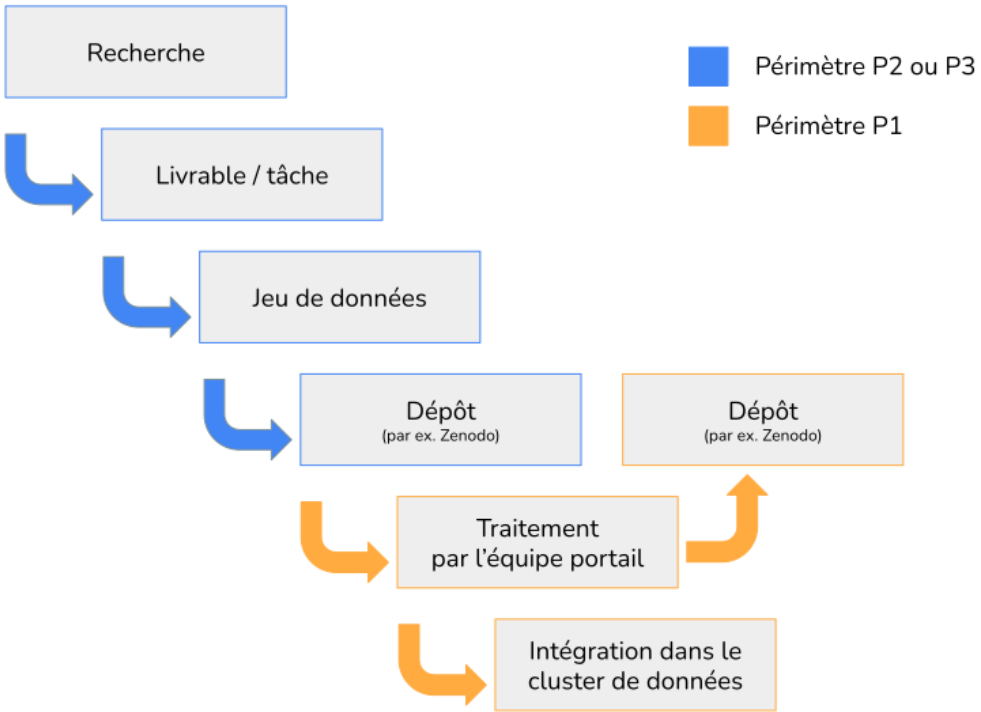

Figure 2: Chaîne de production - traitement des jeux de données

Chaque équipe peut définir le type d'accès qu'elle réserve à ses jeux de données :

- Accès limité uniquement à l'équipe portail ;
- Accès limité uniquement aux équipes partenaires de **Biblissima+** ;
- Accès ouvert à tout le monde (en spécifiant les types de licence pour la réutilisation des données) ;

Quel que soit le type d'accès privilégié, chaque équipe devra également préciser si elle est ou non d'accord pour que l'équipe portail mette à disposition les fichiers enrichis des jeux pour de données versés dans le portail **Biblissima+**. Ces conditions de mise à disposition – limitée à l'équipe partenaire ou à l'ensemble des équipes de partenaires de Biblissima+ ou libre accès pour tous… – sont à indiquer explicitement dans les fichier fichiers README et LICENCE à joindre au dépôt.

L'équipe Portail récupère le jeu de données à traiter depuis le dépôt et procède aux normalisations, alignements et enrichissements décrits plus bas.

6.7 8

Les clusters jouent également un rôle clé pour l'accompagnement et le suivi de la gestion des données. La préparation des agrégations à déposer pour le partage et l'archivage peut bénéficier des réflexions collectives et de mise en commun d'outils ou de workflows éprouvés. Des espaces serveurs partagés fondés sur l'outil « Sharedocs » de l'infrastructure de recherche Huma-Num seront mis à la disposition des clusters pour préparer, documenter et tester les jeux de données préalablement à leur dépôt.

Les droits d'accès seront gérés de manière autonome par les responsables de cluster ou les référents données.

L'équipe portail peut être sollicitée pour prodiguer des conseils ou vérifier que toutes les informations utiles pour l'intégration du jeu de données au sein du portail après dépôt sont présentes et valides.

6.8 9

Le choix de l'entrepôt pour l'archivage et le partage est libre.

En ce qui concerne les codes sources, plusieurs stratégies utilisant les outils de gestion de code open source communément utilisés (Gitlab, Github) en combinaison avec la plateforme d'archivage *Software Heritage* sont définies dans la partie du PGD consacrée au périmètre P1. La stratégie de documentation et de diffusion du jeu de données peut en effet varier selon l'importance scientifique ou patrimoniale des contribution ou le niveau de visibilité souahaitée.

Les PGDs particuliers des codes source produits dans les Périmètres P2 et P3 peuvent s'inspirer de ces stratégies ou s'y référer si cela paraît utile aux responsables scientifiques et techniques.

6.9 10

Zenodo est l'entrepôt principal recommandé pour la diffusion et l'archivage des sets de données.

Il est conseillé de créer un dépôt pour chaque version majeure du produit de recherche. Les jeux de données devront être déposés dans l'une des communautés Zenodo créée pour chaque cluster à cet effet. La curation de la communauté principale https://zenodo.org/communities/biblissima/ est assurée par l'équipe portail. Si les équipes partenaires disposent d'une communauté Zenodo en propre, elles peuvent bien entendu y associer ces dépôts.

6.10 11

En ce qui concerne les dépôts dans d'autres plateformes, ceux-ci sont signalés dans un inventaire géré par le cluster (par exemple sur un cloud collaboratif Sharedocs de l'infrastructure de recherche Huma-Num mis à disposition du cluster).

6.11 12

Lors des journées annuelles des clusters, un point est fait sur les dépôts. Il est suggéré d'organiser des sessions collectives à cette occasion pour tester en groupe la fiabilité des données déposées ou contrôler collectivement la clarté et l'intelligibilité des métadonnées, et de la documentation.

---

1. Document rédigé par les équipes partenaires qui présente en détail l'infrastructure numérique envisagée (téléchargeable depuis la page : https://projet.biblissima.fr/fr/projet/presentation) ↩
2. Cf. https://www.inist.fr/services/accompagner/webinaire/outil-dmp-opidor-modele-de-plan-de-gestion-de-donnees-structure/ ↩
3. Cf. https://argos.openaire.eu/ ↩
4. Le format commun est disponible sur GitHub : https://github.com/RDA-DMP-Common/RDA-DMP-Common-Standard ↩
5. Voir https://www.enseignementsup-recherche.gouv.fr/fr/le-plan-national-pour-la-science-ouverte-les-resultats-de-la-recherche-scientifique-ouverts-tous-49241 ↩
6. https://listes.campus-condorcet.fr/sympa/info/biblissima-donnees ↩
7. Ces recommandations s'inspirent des règles de dépôts des logiciels et codes sources mis en place dans la collaboration Archive HAL/Archive Software Heritage. Voir la page de documentation https://doc.archives-ouvertes.fr/deposer/deposer-le-code-source/ à ce propos. ↩
8. Voir https://www.ionos.fr/digitalguide/sites-internet/developpement-web/fichier-readme/ ↩

7 Exigences minimales

Les jeux de données produits dans le cadre de Biblissima+ susceptible d'être déposés dans une plateforme de partage ou d'archivage de données suivent les consignes de gestion et de préparation des dépôts qui sont décrites ci-après.

Il est recommandé aux équipes partenaires de s'appuyer sur ces dispositions pour rédiger les PGDs particuliers de livrables.

7.1 Nommage des fichiers

Le nommage cohérent et signifiant des noms de fichiers facilite leur classement et permet d'appréhender leur contenu sans avoir à les ouvrir.

Les bonnes pratiques recommandées sont :

- D'éviter les noms trop longs (tout en restant descriptif et clair) ;
- D'éviter les espaces (en utilisant les tirets - et _ comme séparateurs) ;
- D'éviter les caractères non alphanumériques (notamment : & / + > : ? % ( ]) ;
- De normaliser les dates dans le format recommandé par la norme internationale ISO 8601 : YYYY-MM-DD (year-month-day ou année-mois-jour) ;
- D'indiquer la version .

Il est demandé de suivre le schéma de nommage ci-dessous, en particulier pour les jeux de données ayant vocation à être traités par l'équipe portail :

- codeDuLivrable[1]_initiales[2]_dataset_version_dateDeDépôt

Un exemple de nom de jeu de données en suivant ce schéma pourrait être :

- VB_67_CNRS_LM_reperageIntextualite_V1_2027-12-01

7.2 Préparation d'un dépôt

Un espace collaboratif **Sharedocs Huma-Num** sera ouvert pour chaque cluster qui le demande. L'usage d'un tel espace n'est pas obligatoire, mais il permet de travailler collectivement sur la préparation de sets de fichiers à déposer, pendant une période transitoire. Ils n'ont en effet pas vocation à assurer un stockage des données sur une longue durée. **Sharedocs** offre un espace sécurisé pour rassembler, documenter, tester et compléter les ensembles constitués spécifiquement pour les dépôts. Ces espaces peuvent être ouverts à des tiers.

Il est recommandé d'organiser les espaces des clusters sur le même modèle de structuration afin de faciliter les échanges avec l'équipe portail ou entre clusters.

```
Cluster-X
 |__ 1_depots_en_cours
 |__ 2_autres_activites
 |__ 3_ressources
 |__ 4_archives
```

Le répertoire « Ressources » permettra de partager des modèles, des gabarits de fichiers (README, LICENSES, dictionnaires de données, etc.) partageables pour les différents projets et livrables rattachés au cluster.

7.3 Description d'un jeu de données

- Caractérisation des données (types, provenance, formats et standards) ;
- Origine et finalité ;
- Périmètre d'usage (nature, étendue…) ;
- Lien avec des publications scientifiques de type communication, article, chapitre d'ouvrage, ouvrage ou datapaper ;
- Potentiel d'intégration dans d'autres projets ou outils et de réutilisations en général.

7.4 Standards de données et de métadonnées

Citer les standards de données et de métadonnées utilisés.

Le cas échéant, expliquer l'absence de recours à des standards.

7.5 Partage de données

Indiquer comment les données seront partagées :

- Comment est organisé l'accès (plateforme, protocole) ;
- Périodes d'accès restreint avant diffusion ouverte (le cas échéant) ;
- Mécanismes de dissémination ;
- Outils nécessaires à l'exploitation des données (le cas échéant) ;
- Désignation de la plateforme de dépôt.

Si le jeu de données n'est pas partagé, en expliquer les raisons (charte éthique, réglementation concernant la présence de données personnelles, propriété intellectuelle ou commerciale, données sensibles, confidentialité ou sécurité).

7.6 Archivage et préservation

Indiquer comment les données seront archivées et préservées à la fin du projet.

Si des procédures d'archivage à long terme sont mises en place (par exemple dans le cadre d'une convention avec **Huma-Num** et le **CINES**), spécifier la durée pendant laquelle les données devront être préservées, avec des indications sur les volumes à traiter et la manière dont les coûts seront pris en charge.

7.7 Publication des PGDs particuliers

Il est demandé de déposer les PGDs particuliers dans les communautés Zenodo de **Biblissima+** (espace général et du cluster correspondant).

Il est recommandé de rendre le document public et de le mettre tous les deux ans, un mois avant la date de rendu du PGD principal à l'**ANR** avec possibilité d'accès en lecture pour l'équipe technique et le bureau exécutif.

---

1. Pour la référence aux livrables, voir la table de référence dans l'annexe ↩
2. Initiale du créateur du fichier ou du responsable technique et scientifique. ↩

8 Pratiques souhaitées

Il est demandé à tous les participants au projet de respecter collectivement et individuellement les pratiques suivantes :

8.1 1

Créer son identifiant chercheur ORCID ID (http://orcid.org) et le lier à son compte dans l'archive ouverte HAL.

2

Déposer les publications scientifiques (texte intégral) issues du projet dans une archive ouverte, soit directement dans HAL soit par l'intermédiaire d'une archive institutionnelle locale, dans les conditions de l'article 30 de la loi « Pour une République numérique » conformément à la convention attributive d'aide signée avec l'ANR le 26 octobre 2021. Il est recommandé aux auteurs d'utiliser des licences Creative Commons chaque fois que cela est possible et de suivre la stratégie de non cession des droits [1]

- Pour les articles dans des revues, déposer dès la parution le fichier du texte intégral dans HAL avec un embargo d'un an maximum après parution, dans la version validée pour publication ou dans la version avec la mise en page de l'éditeur si celui-ci l'autorise.
- Pour les ouvrages et chapitres d'ouvrages qui ne sont pas couverts par la loi « Pour une République numérique », il est conseillé de négocier avec les éditeurs afin d'insérer dans le contrat d'édition une clause autorisant la possibilité de diffusion en accès ouvert.
- Respecter les consignes de signature de l'établissement de référence et utiliser des identifiants individuels tels que ORCID idHAL qui facilitent l'identification des auteurs et de leurs publications.

8.2 3

Mentionner le soutien apporté par l'ANR au titre du programme d'Investissements d'avenir, en indiquant le numéro de la Convention, dans leurs propres actions de communication sur le Projet « Biblissima+ » (ANR-21-ESRE-0005), ses résultats et dans ses publications, afin qu'elles puissent faire partie du reporting et être prises en compte par les évaluateurs.

Par exemple : « Ce travail a été réalisé grâce à une aide (ou : la consultation) de Biblissima+, qui bénéficie d'une aide de l'Etat gérée par l'ANR au titre du Programme d'investissements d'avenir intégré à France 2030, portant la référence ANR-21-ESRE-0005. ».

8.3 4

Veiller à afficher sur tous les supports de communication orale, les communications par voie d'affiche, les sites internet, etc., le logo **Biblissima+** portant le logo « France 2030 » disponible via ce lien : https://projet.biblissima.fr/fr/logos. Le logo ainsi que les mentions de l'ÉquipEx doivent pointer sur la page racine du projet : https://biblissima.fr.

8.4 5

Participer aux activités des consortium liées à la gestion des données et promouvoir la rédaction de guides de bonnes pratiques, de protocoles ou d'outils pour rendre plus faciles les actions de curation à effectuer, partager des modèles de fichiers README, d'exemples de fiches de métadonnées, de scripts de préparation de données…

8.5 6

S'assurer que les modalités de curation des données, de diffusion et d'archivage pour l'après projet suivent les recommandations établies par le plan de gestion des données.

8.6 7

Porter le PGD à la connaissance de tout nouvel arrivant dans le projet, notamment les personnels recrutés ou les prestataires de services.

8.7 8

Anticiper dans la mesure du possible le recours à l'assistance ou à l'expertise de l'équipe portail.

En ce qui concerne les jeux de données produits dans le cadre de Biblissima 1 (2012-2021) nécessitant une première intégration ou une mise à jour, le responsable scientifique et technique du livrable prend l'initiative de solliciter le concours de l'équipe portail. Les priorités et le calendrier sont définis par l'équipe portail sous la responsabilité du bureau exécutif.

---

1. Pour plus d'informations sur la manière d'appliquer cette stratégie, voir le guide du site "Ouvrir la Science" destiné chercheuses et aux chercheurs https://www.ouvrirlascience.fr/mettre-en-oeuvre-la-strategie-de-non-cession-des-droits-sur-les-publications-scientifiques/ ↩

9 Enjeux des dépôts et identifiants pérennes pour la citation

Les politiques d'incitation au développement de la science ouverte insistent sur l'idée que les jeux de données et les logiciels doivent désormais être considérés comme des produits de recherche légitimes et citables.

Les nouvelles pratiques de citation au sein des publications les incluent dans la liste complète des références, au même titre que les autres résultats de recherche (articles, livres, thèses…). Les organismes et les établissements de recherche sont encouragés à les prendre en compte dans l'évaluation des carrières. Ces productions sont également susceptibles d'être mis en avant dans les réponses aux appels à projets[1].

Étant donné la dimension centrale de l'innovation numérique dans le programme de travail de l'ÉquipEx **Biblissima+**, le dépôt des données et des codes sources représente une dimension importante de sa visibilité et de son évaluation. De même que le partage des données liées aux publications, ces pratiques peuvent également représenter un avantage significatif pour la reconnaissance du travail réalisé par les chercheurs post-doctorants et les ingénieurs qui seront recrutés via l'aide financière obtenue dans ce cadre.

La publication de data papers (publications scientifiques décrivant des jeux de données et leur contexte de production afin de faciliter leur réutilisation) est ainsi fortement recommandée pour accompagner certains dépôts les plus susceptibles d'intéresser la communauté scientifique.

Dans tous les cas, les identifiants pérennes (DOI, SWHID) obtenus lors des opérations de dépôt permettent de construire des citations fiables, durables et donnant directement accès par un lien aux éléments rassemblés[2].

---

1. Voir le guide *Partager les données liées aux publications scientifiques* ↩

2. Voir notamment le document : Féret, Romain, Bracco, Laetitia, Cheviron, Stéphanie, Lehoux, Elise, Arènes, Cécile, & Li, Ling. (2020). Améliorer son projet ANR grâce à la Science Ouverte (Version 2). Zenodo. https://doi.org/10.5281/zenodo.3769954 ↩

10 Responsabilités et ressources

Le paragraphe ci-dessous décrit les responsabilités pour les périmètres P2 et P3. En ce qui concerne le périmètre P1, celles-ci sont décrites dans la partie « PGD détaillé du périmètre P1 ».

10.1 Responsabilités pour les périmètres P2 et P3

Activité	Responsabilités fonctionnelles
Saisie des données	Responsables scientifiques et techniques de livrables
Production des métadonnées	Responsables scientifiques et techniques de livrables
Qualité des métadonnées	Responsables scientifiques et techniques de livrables
Qualité des données	Responsables scientifiques et techniques de livrables
Stockage et sauvegarde	Responsables scientifiques et techniques de livrables
Partage et archivage des données	Responsables scientifiques et techniques de livrables
Rédaction et mise à jour du PGD de livrable	Responsables scientifiques et techniques de livrables
Suivi des dépôts et de la mise à jour du PGD de livrable	Responsables de cluster avec l'assistance du référent données correspondant.
Validation du PGD de livrable	Responsable d'unité de l'Équipe partenaire au sein de laquelle le livrable est produit.

10.2 Précisions sur le rôle de référent données au sein des clusters

Les « référents données » ont pour mission principale de participer au groupe transversal fédéré autour de la liste de discussion *biblissima-donnees*. Ils assurent un rôle de relais entre les membres du cluster, l'équipe portail et le bureau exécutif en lien avec les responsables de cluster. Ils créent les espaces Zenodo et sont ainsi informés de la mise à disposition de nouveaux jeux de données.

Ils peuvent également initier ou coordonner des réflexions sur l'harmonisation des pratiques et le développement de méthodologies ou d'outils mutualisés pour lesquels un double regard technique et scientifique est nécessaire.

Ils jouent également un rôle d'orientation et de conseil « de premier niveau » en ce qui concerne l'articulation des PGDs particuliers avec les recommandations touchant les périmètres P2 et P3 au sein du PGD principal.

#	Cluster	Référents données
Cluster 1	Acquisition des corpus de sources interopérables	Mathieu STOLL (SIAF)
Cluster 2	Prise en compte et cherchabilité des données d'analyse des matériaux	Anne MICHELIN (CRC)
Cluster 3	Intelligence artificielle, reconnaissance de formes et d'écritures manuscrites	Dominique STUTZMANN (IRHT)
Cluster 4	Traitement approfondi des systèmes graphiques et analyse des documents	Peter STOKES (AOrOC)
Cluster 5a	TEI et épigraphie, de l'Antiquité à l'époque moderne	Michèle BRUNET (HiSoMA)
Cluster 5b	Édition de sources en TEI	Stéphane LECOUTEUX (MRSH)

#	Cluster	Référents données
Cluster 6	Les défis du patrimoine musical	David FIALA (CESR)
Cluster 7	Interopérabilité et analyse des textes	Jean-Baptiste CAMPS (CJM)

III. Vue d'ensemble des jeux de données

11 Vue d’ensemble des données

Les 6 tableaux présentés ci-après récapitulent l’ensemble des livrables proposés. Les périmètres concernés sont indiqués entre parenthèses dans le sommaire ci-dessous.

12 1 – Infrastructure numérique

Produit de recherche	Description	Nature des données	Formats / standards	Volumétrie	Politique de partage	Politique de conservation à long terme	Actions de Fairisation à mener
Cluster de données (Portail)	Données importées dans le Portail (base Postgresql), publiées via l'application web CubicWeb (Python)	Données textuelles formalisées	MARC, XML, SQL, RDF	776 231 entités (25/10/2023) , pour 1,5 Go de données XML	Publication des données via le Portail, exposition de données dans le Web sémantique. Partage via les dépôts des jeux de données au format pivot et enrichis.	Archivage en fin de projet d'un export des données au format RDF sur Zenodo. Les ressources issues des partenaires sont déposées de manière autonome par leur producteurs à chaque version majeure (cf. processus d'intégration de sources de données dans le portail).	Préparer l'archivage final à la fin du projet en même temps que les spécifications des développements technologiques de l'infrastructure héritée de Biblissima 1. Sur les notices de ressources afficher les DOI des dépôts des jeux de données.
Cluster de données (IIIF-Collections)	Données importées dans IIIF-Collections (ElasticSearch), publiées via une application PHP	Données textuelles formalisées	CSV, JSON	89 613 items (25/10/2023)	Publication des données via le site IIIF-Collections	Les données vont être transformées en XML et versées dans le Portail (cf. ligne ci-dessus)	N / A
Interfaces et applications web	Moteurs d'indexation et interfaces de recherche et de visualisation	Codes informatiques	PHP, Python, JSON, Javascript	Portail : ~500Mo de code + ~5.5Go de caches et tests	Via Github / Gitlab et Zenodo pour les versions majeures	Moissonnage par l'archive pérenne de logiciels Software	N / A

Produit de recherche	Description	Nature des données	Formats / standards	Volumétrie	Politique de partage	Politique de conservation à long terme	Actions de Fairisation à mener
	des données (Portail et IIIF-Collections), développées en interne et en lien avec des prestataires. Intégration de la recherche sur les matériaux dans l'interface du portail (avec CRC)			d'import – IIF Collections (app web) : ~1Go		Heritage. Les logiciels ou modules dotés d'un potentiel de réutilisation dans la communauté feront l'objet d'un dépôt avec métadonnées modéré via la voie couplée HAL + Software heritage	
Visualiseur d'images Mirador	Version packagée du visualiseur (avec des plugins) pour le Portail Biblissima	Codes informatiques	Javascript, IIIF	~13 Mo	Github / Gitlab	N / A	N / A
Plateforme des référentiels et ses API data.bilissima.fr	Plateforme d'édition et d'exposition des référentiels d'autorités	Codes informatiques	Wikibase, PHP	~1.1Go	Utilisationd de la technologie Wikibase afin de créer un "hub" d'identifiants et de données structurées, accessibles, interopérables et réutilisables. Le hub	N / A	Non

Produit de recherche	Description	Nature des données	Formats / standards	Volumétrie	Politique de partage	Politique de conservation à long terme	Actions de Fairisation à mener
					donne les PIDs des entités (URIs déréférençables) et de leur documentation pour les utilisateurs et via une API web et un Sparql endpoint pour l'accès distant à des programmes informatiques.		
Référentiels d'autorité et thésaurus iconographique	Vocabulaires contrôlés pour lier entre elles les ressources du portail intégrées au cluster de données	Données textuelles formalisées	RDF, Json	5 Go pour l'ensemble	Dépôt des versions majeures de dumps RDF par référentiel.	Archivage en fin de projet d'un export des données au format RDF sur Zenodo.	Rédaction d'un data paper par référentiel après dépôt.
Études, cahiers des charges, spécifications, documentation des processus, etc.	Documentation interne des développements informatiques	Données textuelles	.docx, .pdf, .md	N / A	Non partagé a priori, peut être inclus dans les dépôts des codes sources si utile à l'intelligibilité des données.	Non sauf si intégré à la documentation d'un dépôt archivé.	N / A

13 2 – Outils de traitements de données du cluster de données

Produit de recherche	Description	Nature des données	Formats / standards	Volumétrie	Politique de partage	Politique de conservation à long terme	Actions restant à mener pour B1
Format Pivot	Format d'entrée des données (modélisé avec alignements partiels sur la TEI, CiDOC-CRM et FRBRoo)	Données textuelles formalisées	DTD XML	1 fichier XML	Présenté sur doc.biblissima.fr avec la documentation du processus d'harmonisation des données et d'intégration dans le cluster de données (dite "Vademecum") et diffusé via Github / Gitlab.	HAL + Software Heritage	Dépôt HAL
Scripts de conversion et de traitements	Scripts spécifiques pour chaque source de données à intégrer (moissonnage, transformation, import)	Codes informatiques	PHP, Python	1 fichier par version de source / env. 40 unités traitées – IIIF Collections (scripts + données) : ~1.1Go – Traitement des sources de données de B+ : cf tableaux suivants	Stockage sur les serveurs du CC (Seafile), Biblissima 1 : non diffusé / Biblissima+ : Gitlab	Gitlab + Software Heritage	Aucune pour B1 (Intégré à la chaîne de traitement pour B+)
Webservice de réconciliation et d'alignement de données	Service permettant à tout projet d'aligner ses données avec les	Codes informatiques	JSON (Wikibase manifest)	1 fichier manifest	Publié sur la plateforme publique d'Open Refine sur Github sous	Github + Software Heritage	Vérifier moissonnage auto dans Software Heritage

Produit de recherche	Description	Nature des données	Formats / standards	Volumétrie	Politique de partage	Politique de conservation à long terme	Actions restant à mener pour B1
pour OpenRefine	référentiels de data.biblissima.fr dans l'outil libre OpenRefine ou autre				forme de wikibase-manifest		
Mécanismes et protocoles de mise à jour des sources intégrées au portail	Développements pour l'enrichissement et l'évolution de l'infrastructure portail	Codes informatiques	Selon les besoins et spécifications	quelques Mo/Go	Hébergement sur l'entrepôt git du prestataire Logilab (Mercurial) avec clone sur les serveurs de Biblissima+	N / A	Sans objet

14 3 – Ressources référencées par le portail ou intégrées au référentiel d’autorités

Type de données brutes fournies	Nom	Équipe	Type de ressource selon catégories du Portail	Utilisation native des référentiels Biblissima	Suivi intégration	Licence de diffusion du partenaire pour la base source	Politique de conservation à long terme du partenaire pour la base source
Export SQL	Esprit des livres	EnC	Catalogue ou répertoire	NON	B1, intégré	Propriété intellectuelle ENC	non connu
Export SQL	Jonas	IRHT-CNRS	Base de données scientifiques	NON	B1, intégré	Propriété intellectuelle IRHT	non connu
Export CSV	Comparatio	IRHT-CNRS	Base de données scientifiques	NON	B1, intégré	Propriété intellectuelle IRHT	non connu
Export Marc-XML	CR2I	CESR	Catalogue ou répertoire	NON	B1, intégré	non connu	non connu
Export Marc-XML	Wellcome collection	Hors partenariat	Catalogue ou répertoire	NON	B1, intégré	CC BY 4.0	non connu
Export EAD-XML	BnF Archives et Manuscrits	BnF	Catalogue ou répertoire	NON	B1, intégré	Licence ouverte	non connu
Export TEI	Reliures.bnf.fr	Bnf	Base de données scientifiques	NON	B1, non intégré	Licence ouverte	non connu
Export TEI	Miroir des classiques	EnC	Catalogue ou répertoire	NON	B1, non intégré	CC BY NC ND 2.0	non connu
Export XML mixte (.csv, dat…)	Mandragore	BnF	Catalogue ou répertoire	NON	B1, intégré	Licence ouverte	non connu
XML Pivot exporté	Bibale	IRHT-CNRS	Base de données scientifiques	NON	B1, intégré	CC BY NC	non connu

Type de données brutes fournies	Nom	Équipe	Type de ressource selon catégories du Portail	Utilisation native des référentiels Biblissima	Suivi intégration	Licence de diffusion du partenaire pour la base source	Politique de conservation à long terme du partenaire pour la base source
XML Pivot exporté	Pinakes	IRHT-CNRS	Base de données scientifiques	NON	B1, intégré	Propriété intellectuelle IRHT	Infrastructure IRHT
XML Pivot exporté	Bibliothèques françoises	CESR	Corpus ou édition de source	NON	B1, intégré	CC BY-NC-SA	non connu
XML Pivot exporté	RegeCart	IRHT-CNRS	Base de données scientifiques	NON	B1, intégré	Propriété intellectuelle IRHT	non connu
XML Pivot dynamique	Manuscripta Medica	SAPRAT-EPHE, CIHAM	Base de données scientifiques	NON	B1, intégré	non connu	non connu
XML Pivot dynamique	Initiale	IRHT-CNRS	Catalogue ou répertoire	NON	B1, intégré	CC BY-NC 3.0	Infrastructure IRHT
XML Pivot dynamique	Books within books	SAPRAT-EPHE	Base de données scientifiques	NON	B1, non intégré	non connu	non connu
OAI-PMH (METS)	Heidelberger historische Bestände	Hors partenariat	Corpus ou édition de source	NON	B1, intégré	Public Domain Mark	non connu
OAI-PMH (TEI)	Volumes de la série « Documents, études et répertoires de l'Institut de Recherche et d'Histoire des Textes (DER)	IRHT-CNRS, Persée	Catalogue ou répertoire	NON	B1, intégré	CC BY NC SA	Archivage pérenne au CINES de la plateforme Persée dans son ensemble

Type de données brutes fournies	Nom	Équipe	Type de ressource selon catégories du Portail	Utilisation native des référentiels Biblissima	Suivi intégration	Licence de diffusion du partenaire pour la base source	Politique de conservation à long terme du partenaire pour la base source
OAI-PMH (DC)	Médiathèque de Moulins, Mss et imprimés	Hors partenariat	Catalogue ou répertoire	NON	B1, intégré	non connu	non connu
Manifestes IIIF + MARC-XML	Universiteits bibliotheek Gent - Digitale bibliotheek	Hors partenariat	Catalogue ou répertoire	NON	B+, intégré	Licences variables selon les documents (CC, RightsStatements)	non connu
Manifestes IIIF	Leiden University Libraries Digital Collections	Hors partenariat	Catalogue ou répertoire	NON	B+, intégré	CC BY	non connu
Manifestes IIIF	Bibliothèques Virtuelles Humanistes. Fac-similés	CESR	Catalogue ou répertoire	NON	B+, intégré	non connu	non connu
Manifestes IIIF	Mmmonk	Hors partenariat	Catalogue ou répertoire	NON	B+, intégré	CC BY, CC BY NC	non connu
SRU-SRW (DC)	PaGella (Bibliothèque municipale de Grenoble). Mss et imprimés anciens numérisés	Hors partenariat	Catalogue ou répertoire	NON	B+, intégré	Domaine public	non connu

Type de données brutes fournies	Nom	Équipe	Type de ressource selon catégories du Portail	Utilisation native des référentiels Biblissima	Suivi intégration	Licence de diffusion du partenaire pour la base source	Politique de conservation à long terme du partenaire pour la base source
SRU-SRW (DC) + RDF-XML	Rotomagus (Bibliothèque municipale de Rouen). Mss et imprimés anciens numérisés	Hors partenariat	Catalogue ou répertoire	NON	B+, intégré	Domaine public	non connu
Manifestes IIIF	Bibliothèque numérique CBMA	Hors partenariat	Catalogue ou répertoire	NON	B+, intégré	RightsStatements “In Copyright”	non connu
SRU-SRW (DC) + RDF-XML	Mémonum (Bibliothèque municipale de Montpellier). Mss et imprimés anciens numérisés	Hors partenariat	Catalogue ou répertoire	NON	B+, intégré	Domaine public	non connu
Manifestes IIIF	Biblioteca Nacional de Portugal. Mss et imprimés anciens numérisés	Hors partenariat	Catalogue ou répertoire	NON	B+, intégré	Domaine public	non connu
Manifestes IIIF + MODS-XML	Parker Library (Cambridge). Mss numérisés	Hors partenariat	Catalogue ou répertoire	NON	B+, intégré	CC BY NC	non connu
Manifestes IIIF	Bibliothèque Mazarine.	Hors partenariat	Catalogue ou répertoire	NON	B+, intégré	non connu	non connu

Type de données brutes fournies	Nom	Équipe	Type de ressource selon catégories du Portail	Utilisation native des référentiels Biblissima	Suivi intégration	Licence de diffusion du partenaire pour la base source	Politique de conservation à long terme du partenaire pour la base source
	Mss et imprimés anciens numérisés						
Manifestes IIIF	Bibliothèque de Saint-Omer. Mss et imprimés anciens numérisés	Hors partenariat	Catalogue ou répertoire	NON	B+, intégré	non connu	non connu
OAI-PMH (METS-MODS)	Staats- und Universitätsbibliothek Bremen. Digitale Sammlungen	Hors partenariat	Catalogue ou répertoire	NON	B+, intégré	Public Domain Mark	non connu
à définir	BBMN Montfaucon	IRHT, MRSH	Corpus ou édition de source	NON	B1, non intégré	non connu	non connu
à définir	Collecta	IRHT, BnF	Catalogue ou répertoire	NON	B1, non intégré	non connu	non connu
à définir	E-ktobe	IRHT	Catalogue ou répertoire	NON	B1, non intégré	Propriété intellectuelle IRHT	Infrastructure IRHT
export TEI	Gloss-e	IRHT, LEM	Corpus ou édition de source	NON	B1, non intégré	non connu	non connu
à définir	Sermones.net	IRHT, CIHAM	Corpus ou édition de source	NON	B1, non intégré	non connu	non connu

Type de données brutes fournies	Nom	Équipe	Type de ressource selon catégories du Portail	Utilisation native des référentiels Biblissima	Suivi intégration	Licence de diffusion du partenaire pour la base source	Politique de conservation à long terme du partenaire pour la base source
export TEI	SourcEncyMe	IRHT	Corpus ou édition de source	NON	B1, non intégré	Propriété intellectuelle IRHT	Serveurs IRHT (Orléans)
à définir	Sanderus electronicus	IRHT, MRSH	Corpus ou édition de source	NON	B1, non intégré	non connu	non connu
à définir	Inventaires manuscrits grecs	IRHT	Catalogue ou répertoire	NON	B1, non intégré	Propriété intellectuelle IRHT	Infrastructure IRHT
à définir	Données du CCFr	Bnf	référentiel pour data.biblissima.fr	NON	B+, non intégré	Licence ouverte	voir politique BnF
			Catalogue ou répertoire				
à définir	Données de Persée	Persée	référentiel pour data.biblissima.fr	OUI	B+, non intégré	CC BY-NC-SA 3.0	voir politique Persée
			Catalogue ou répertoire				
à définir	Données d'ISMI	IRHT, B+	référentiel pour data.biblissima.fr	NON	B+, non intégré	Licence ouverte	Infrastructure IRHT
à définir	Référentiels de noms de lieux et de personnes dans les	IRHT	référentiel pour data.biblissima.fr	NON	B+, non intégré	CC BY	Via Nakala

Type de données brutes fournies	Nom	Équipe	Type de ressource selon catégories du Portail	Utilisation native des référentiels Biblissima	Suivi intégration	Licence de diffusion du partenaire pour la base source	Politique de conservation à long terme du partenaire pour la base source
	cartulaires médiévaux						
via une API	Référentiels pour les manuscrits de l'Orient chrétien et byzantin (grec, syriaque, arabe)	IRHT, CJM	référentiel pour data.biblissima.fr	NON	B+, non intégré	Propriété intellectuelle IRHT	Infrastructure IRHT
à définir	Thesaurus numismatique	CRAHAM	référentiel pour data.biblissima.fr	NON	B+, non intégré	CC BY	Sans objet
à définir	Institutions ecclésiastiques anciennes (via travaux du consortium COSME 2)	IRHT	référentiel pour data.biblissima.fr	NON	B+, non intégré	non connu	non connu
à définir	Référentiels valeurs et mesures	CIHAM	référentiel pour data.biblissima.fr	NON	B+, non intégré	non connu	non connu
via une API	Données de Biblindex en diffusion ouverte (12 Bibles, métadonnées de 3000 oeuvres,	HISOMA	Corpus ou édition de source	OUI	B+, non intégré	CC BY	sans objet

Type de données brutes fournies	Nom	Équipe	Type de ressource selon catégories du Portail	Utilisation native des référentiels Biblissima	Suivi intégration	Licence de diffusion du partenaire pour la base source	Politique de conservation à long terme du partenaire pour la base source
	autres métadonnées						
à définir	Répertoire de filigranes	IRHT	Catalogue ou répertoire	NON	B+, non intégré	non connu	non connu
API	Corpus des inscriptions de la France médiévale, vol. 1-25 (extraction OCR des anciens volumes numérisés dans Persée, enrichissement et encodage XML-TEI des notices, intégration dans la base de données Titulus)	CESCM	Corpus ou édition de source	non connu	B+, non intégré	Licence ouverte	CINES
API	Corpus des inscriptions de la France médiévale, vol. 26 et HS 1-3 (encodage XML-TEI des volumes à paraître, intégration des notices	CESCM	Corpus ou édition de source	non connu	B+, non intégré	Licence ouverte	CINES

Type de données brutes fournies	Nom	Équipe	Type de ressource selon catégories du Portail	Utilisation native des référentiels Biblissima	Suivi intégration	Licence de diffusion du partenaire pour la base source	Politique de conservation à long terme du partenaire pour la base source
	dans la base de données Titulus)						
à définir	CLEM Carmina Latina Epigraphica Moderna	CJM	Corpus ou édition de source	non connu	B+, non intégré	Licence CC ?	à définir
à définir	Edition des Gloses	CIHAM - IRHT	Corpus ou édition de source	OUI	B+, non intégré	CC BY ou Etalab 2.0	Zenodo
à définir	Édition de sources de types différents (littéraires, encyclopédiques, diplomatiques, sources de la pratique)	CRAHAM	Corpus ou édition de source	OUI	B+, non intégré	Nakala, licence CC BY	non concerné
API	Thesauri et autorités en lien avec les projets d'édition	PDN de la MRSH de Caen, CRAHAM	référentiel pour data.biblissima.fr	OUI	B+, non intégré	Nakala, licence CC BY	non concerné
API	Thesaurus Ichtya (noms de poissons et référentiels aquatiques)	PDN de la MRSH de Caen, CRAHAM	référentiel pour data.biblissima.fr	OUI	B+, non intégré	Nakala, licence CC BY	non concerné

Type de données brutes fournies	Nom	Équipe	Type de ressource selon catégories du Portail	Utilisation native des référentiels Biblissima	Suivi intégration	Licence de diffusion du partenaire pour la base source	Politique de conservation à long terme du partenaire pour la base source
à définir	Bibliothèque Ichtya	PDN de la MRSH de Caen, CRAHAM	Corpus ou édition de source	OUI	B+, non intégré	Nakala, licence CC BY	non concerné
à définir	Textes liturgiques à l'usage du Mont Saint-Michel	CRAHAM	Corpus ou édition de source, data.biblissima.fr	OUI	B+, non intégré	Nakala, licence CC BY	non concerné
à définir	Edition des actes de Gautier de Coutances (archevêque de Rouen (1184-1207)	CRAHAM	Corpus ou édition de source, data.biblissima.fr	OUI	B+, non intégré	Nakala, licence CC BY	non concerné
à définir	Edition des livres 3 et 4 de l'Histoire du grand comte Roger… par Geoffroi Malaterra	CRAHAM	Corpus ou édition de source, data.biblissima.fr	Peut-être	B+, non intégré	Nakala, licence CC BY	non concerné
API	Données Biblindex	HiSoMA	data.biblissima.fr	OUI	B+, non intégré	Zenodo, licence ouverte	non concerné
à définir	Chaînage d'outils d'édition : acquisition de données (VB_35_ENC ,	CJM	data.biblissima.fr	OUI	B+, non intégré	Github, CC BY	Nakala

Type de données brutes fournies	Nom	Équipe	Type de ressource selon catégories du Portail	Utilisation native des référentiels Biblissima	Suivi intégration	Licence de diffusion du partenaire pour la base source	Politique de conservation à long terme du partenaire pour la base source
	VB_37_ENC et PE_18_ENC)						
à définir	Miroir des classiques : éditions partielles de traductions du Corpus juris civilis	CJM	Corpus ou édition de source	NON	B+, non intégré	Github, CC BY	Nakala
à définir	Projet Wala : indexation des sources musicales de l'ouest de la France	IRHT	non connu	non connu	B+, non intégré	non connu	non connu
Export MEI	Edition MEI de partitions musicales (40 recueils musicaux)	CESR	Corpus ou édition de source	non connu	B+, non intégré	non connu	CINES
à définir	Corpus lexical européen (50 M mots de latin médiéval européen 700-1300)	IRHT	Corpus ou édition de source	non connu	B+, non intégré	non connu	non connu
à définir	Corpus lexical Velum	IRHT	Corpus ou édition de source	non connu	B+, non intégré	non connu	non connu

Type de données brutes fournies	Nom	Équipe	Type de ressource selon catégories du Portail	Utilisation native des référentiels Biblissima	Suivi intégration	Licence de diffusion du partenaire pour la base source	Politique de conservation à long terme du partenaire pour la base source
Moissonnage IIIF	Manifests IIIF produits et exposés par le portail FranceArchives	SIAF	Catalogue ou répertoire	nan	B+, non intégré	Licence ouverte	non concerné

15 4 – Chaînes d’outils logiciels

Produit de recherche	Description	Équipe	Nature des données	Formats / standards	Volumétrie	Intégration	Politique de partage	Politique de conservation à long terme
Collatinus	Lemmatiseur et analyseur morphologique de textes, version bureau latins (boîte à outils Baobab)	Développeur open source + Phlam	Codes informatiques	Qt (C++)	non connu	B1, référencé	Installateurs téléchargeables sur le site Baobab + paquet disponible dans les dépôts Debian – Code source disponible sur Github (GNU General Public License v3.0)	Software Heritage
Collatinus-web	Lemmatiseur et analyseur morphologique de textes latins, version web	Développeur open source + Phlam + Equipe Biblissima	Codes informatiques	Qt (C++), PHP, Javascript	non connu	B1, référencé	Démon C++ (partie serveur) disponible dans une branche du dépôt collatinus (cf. ci-dessus) – Application web (partie cliente) intégrée dans un conteneur Jekyll télécharge	Software Heritage

Produit de recherche	Description	Équipe	Nature des données	Formats / standards	Volumétrie	Intégration	Politique de partage	Politique de conservation à long terme
							able via Github	
Eulexis	Logiciel de lemmatisation de textes en grec ancien, version bureau	Phlam	Codes informatiques	Qt (C++)	non connu	B1, référencé	Installateurs téléchargeables sur le site Baobab – Code source disponible sur Github (GNU General Public License v3.0)	Software Heritage
Eulexis-web	Logiciel de lemmatisation de textes en grec ancien, version web	Phlam + Équipe Biblissima	Codes informatiques	PHP, Javascript	non connu	B1, référencé	Application PHP et Javascript, intégrée dans un conteneur Jekyll téléchargeable via Github	Software Heritage
Développement d'Eulexis	Intégration des données et fonctionnalités du lemmatiseur Hisoma dans Eulexis	HiSoMA, Phlam	Données textuelles	.csv	50.000 couples lemmes-formes, et autres enrichissements	B+, à intégrer	Licence ouverte	via Eulexis
Praelector	Assistant de lecture	Développeur open	Codes informatiq	Qt (C++)	5 Mo	B1, référencé	Version à télécharger	Software Heritage

Produit de recherche	Description	Équipe	Nature des données	Formats / standards	Volumétrie	Intégration	Politique de partage	Politique de conservation à long terme
	du latin (version en test)	source	ues				sur le site Biblissima, sources sur Debian Gitlab. Licence GNU GPL v3.	auto
Schémas reliures	Schéma d'encodage TEI formalisé et documenté pour les reliures de livres anciens	BNF - Réserve des livres rares	Données textuelles formalisées	ODD (XML-TEI)	1 fichier source ODD + déclinaisons dans .xsd, .rng, etc.	B1, référencé	Présentation et lien de téléchargement publié sur le site de la BNF et sur le site Biblissima	Utilisation du format ODD (utilisé par le CINES pour traiter la TEI)
Outils d'édition XML	Environnement d'encodage via des interfaces conviviales - PDN Caen et Certic	PDN et Certic (Caen)	Données textuelles formalisées, codes informatiques	XML-TEI, XML-EAD, JAVA	non connu	B1, référencé	Diffusion au téléchargement sur le site des Presses Document numérique de l'université de Caen (PDN). Licences : Cecill (catalogage EAD) GNU GPL v3 (Inventaires anciens en XML-TEI) et	Non connu

Produit de recherche	Description	Équipe	Nature des données	Formats / standards	Volumétrie	Intégration	Politique de partage	Politique de conservation à long terme
							Cecill-C (Pluco)	
Outil Thecae	Application web MaX de publication de la collection La collection Thecae, Corpus d'inventaires anciens de livres manuscrits et imprimés	PDN de Caen	Codes informatiques	XQuery, XML, HTML, CSS, Javascript	non connu	B1, référencé	Non partagé	Non connu
MaX	Moteur d'affichage XML (application web BaseX préconfigurée et personnalisable)	PDN de la MRSH et Certic (Caen)	Codes informatiques	XQuery, XML, HTML, CSS, Javascript	2 Mo	B1, référencé	Diffusion sur la plateforme Gitlab de l'Université de Caen, sous licence Cecill-B	Non connu
Protocoles et outils pour les corpus et éditions XML	Service de partage de textes DTS	CJM	Codes informatiques	XQuery, XML, HTML, CSS, Javascript, JAVA	non connu	B+, non intégré	Diffusion sur Github, licences open source à définir au cas par cas (CC BY ou licence ouverte la	Software Heritage

Produit de recherche	Description	Équipe	Nature des données	Formats / standards	Volumétrie	Intégration	Politique de partage	Politique de conservation à long terme
							plupart du temps)	
Protocoles et outils pour les corpus et éditions XML	Développements pour TEI Publisher	HiSoMA	Codes informatiques	XQuery, XML, HTML, CSS, Javascript, JAVA	non connu	B+, non intégré	Diffusion sur Github, licences open source à définir au cas par cas (CC BY ou licence ouverte la plupart du temps)	Software Heritage
Protocoles et outils pour les corpus et éditions XML	Protocole d'encodage des citations de la Bible	HiSoMA, cluster 5b	Codes informatiques	XQuery, XML, HTML, CSS, Javascript, JAVA	non connu	B+, non intégré	Diffusion sur Github, licences open source à définir au cas par cas (CC BY ou licence ouverte la plupart du temps)	Software Heritage
Protocoles et outils pour les corpus et éditions XML	Nouveaux environnement de balisage et de publication	PDN de la MRSH de Caen, CRAHAM, cluster 5b	Codes informatiques	XQuery, XML, HTML, CSS, Javascript, JAVA	non connu	B+, non intégré	Diffusion sur Github, licences open source à définir au cas par cas (CC BY ou licence ouverte la	Software Heritage

Produit de recherche	Description	Équipe	Nature des données	Formats / standards	Volumétrie	Intégration	Politique de partage	Politique de conservation à long terme
							plupart du temps)	
Protocoles et outils pour les corpus et éditions XML	Développement de configurations types pour le moteur d'affichage Max	IRHT, MRSH de Caen	Codes informatiques	XQuery, XML, HTML, CSS, Javascript, JAVA	non connu	B+, non intégré	Diffusion sur Github, licences open source à définir au cas par cas (CC BY ou licence ouverte la plupart du temps)	Software Heritage
Protocoles et outils pour les corpus et éditions XML	Configuration de pluCo pour Oxygen (manuel)	IRHT cluster 5b	Codes informatiques	XQuery, XML, HTML, CSS, Javascript, JAVA	non connu	B+, non intégré	Diffusion sur Github, licences open source à définir au cas par cas (CC BY ou licence ouverte la plupart du temps)	Software Heritage
Protocoles et outils pour les corpus et éditions XML	Développement de configurations types pour le moteur d'affichage Max	IRHT cluster 5b	Codes informatiques	XQuery, XML, HTML, CSS, Javascript, JAVA	non connu	B+, non intégré	Diffusion sur Github, licences open source à définir au cas par cas (CC BY ou licence ouverte la	Software Heritage

Produit de recherche	Description	Équipe	Nature des données	Formats / standards	Volumétrie	Intégration	Politique de partage	Politique de conservation à long terme
							plupart du temps)	
Protocoles et outils pour les corpus et éditions XML	Développement d'une solution conviviale pour le travail collaboratif dans Oxygen	IRHT cluster 5b	Codes informatiques	XQuery, XML, HTML, CSS, Javascript, JAVA	non connu	B+, non intégré	Diffusion sur Github, licences open source à définir au cas par cas (CC BY ou licence ouverte la plupart du temps)	Software Heritage
Protocoles et outils pour les corpus et éditions XML	Chaînage d'outils d'édition : développement applicatif	CJM	Codes informatiques	XQuery, XML, HTML, CSS, Javascript, JAVA	non connu	B+, non intégré	Diffusion sur Github, licences open source à définir au cas par cas (CC BY ou licence ouverte la plupart du temps)	Software Heritage
Protocoles et outils pour les corpus et éditions XML	Amélioration incrémentielle d'un plugin TEI pour un éditeur XML libre (JEdit) en lien avec	CIHAM	Codes informatiques	XQuery, XML, HTML, CSS, Javascript, JAVA	non connu	B+, non intégré	Diffusion sur Github, licences open source à définir au cas par cas (CC BY ou licence ouverte la	Software Heritage

Produit de recherche	Description	Équipe	Nature des données	Formats / standards	Volumétrie	Intégration	Politique de partage	Politique de conservation à long terme
	le plugin pluCo						plupart du temps)	
Portail du laboratoire d'édition et d'annotation de sources	Espace d'expérimentations et de développement d'interfaces d'encodage et de publication,	PDN de la MRSH de Caen, CRAHAM, cluster 5b	Codes informatiques, données textuelles formalisées	XQuery, XML, HTML, CSS, Javascript, IIIF, DTS	non connu	B+, non intégré	à définir, le travail sur les sources reste travail sur les sources restera protégé par le droit d'auteur	Software Heritage
Portail du laboratoire d'édition et d'annotation de sources	tests sur les sources encodées en XML-TEI et réflexion avec le PDN et les autres partenaires sur l'outillage des sources,	PDN de la MRSH de Caen, CRAHAM, cluster 5b	Codes informatiques, données textuelles formalisées	varia	non connu	N / A	N / A	N / A
Portail du laboratoire d'édition et d'annotation de sources	Réflexions communes sur les méthodologies d'encodage	IRHT	Codes informatiques, données textuelles formalisées	XML/TEI, ODD	non connu	N / A	CC By à définir	non connu
Portail du laboratoire d'édition	Protocole d'encodage des	HiSoMA, cluster 5b	Codes informatiques,	XML/TEI, ODD	non connu	N / A	non connu	non connu

Produit de recherche	Description	Équipe	Nature des données	Formats / standards	Volumétrie	Intégration	Politique de partage	Politique de conservation à long terme
et d'annotation de sources	citations de la Bible		données textuelles formalisées					
Portail du laboratoire d'édition et d'annotation de sources	Schémas documentés	CJM	Codes informatiques, données textuelles formalisées	XML/TEI, ODD	non connu	N / A	non connu	non connu
Développement d'outils innovants pour les recherches de l'IRHT sur les textes latins et français	Classification des éléments graphiques (pages et zones de pages) – Catalogage automatique des manuscrits numérisés : identification des textes issus de HTR par comparaison avec référentiels textuels – Reconnaissance d'entités nommées et alignement	IRHT - TEKLIA	données textuelles formalisées	non connu	non connu	Voir le PGD particulier du livrable	Voir le PGD particulier du livrable	Voir le PGD particulier du livrable

Produit de recherche	Description	Équipe	Nature des données	Formats / standards	Volumétrie	Intégration	Politique de partage	Politique de conservation à long terme
	sur des référentiels							
Développement de TELMA-ANACLET	ANalyse Approfondie de Corpus éLEctroniques Textuels : traitement a posteriori des données par l'utilisateur	IRHT	codes informatiques	CMS ?	non connu	Voir le PGD particulier du livrable	Voir le PGD particulier du livrable	Voir le PGD particulier du livrable
Développement de Kraken	Développement et maintenance de la suite d'outils Kraken (Cluster 3)	AOROC	codes informatiques	voir eScriptorium – Python ? XML ALTO ?	non connu	Voir le PGD particulier du livrable – Archive de modèles sur Zenodo	Voir le PGD particulier du livrable	Voir le PGD particulier du livrable
Reconnaissance automatisée de coins monétaires	Réalisation d'un système automatique de reconnaissance des coins monétaires antiques (cluster X)	AOROC/Ecole des Mines de Paris	codes informatiques	non connu	non connu	Voir le PGD particulier du livrable	Voir le PGD particulier du livrable	Voir le PGD particulier du livrable
Ressources computati	Mise à disposition	CJM	codes informatiq	python	non connu	Voir le PGD	Github, CC By	Zenodo

Produit de recherche	Description	Équipe	Nature des données	Formats / standards	Volumétrie	Intégration	Politique de partage	Politique de conservation à long terme
onnelles pour le traitement automatique des langues historiques à forte variation graphique	d'outils et de modèles - utilisation de l'outil Pie pour entraîner les modèles		ues			particulier du livrable		

16 5 – Autres ressources

Produit de recherche	Description	Équipe	Cluster (volet B)	Nature des données	Formats / standards	Volumétrie	Politique de partage	Politique de conservation à long terme	Actions de Fairisation à mener
Développements réalisés par les équipes partenaires et intégrés à des systèmes d'information existant ou configuration particulières d'outils	Visionneuse IIIF (SIAF)	voir ci-contre	-	codes informatiques	Javascript, IIIF, Drupa, PHP, RDF, Python…	Non connu	Github / Gitlab/ Nakala et diffusion sous licence ouverte – Voir le PGD particulier du livrable	Voir le PGD particulier du livrable	Voir le PGD particulier du livrable – Actualisation du modèle de données du format pivot XML si la base cible est intégrée au portail
Développements réalisés par les équipes partenaires et intégrés à des systèmes d'information existant ou configuration particulières d'outils	Plateforme Savoirs (CRH/EHESS éditions)	voir ci-contre	-	codes informatiques	Javascript, IIIF, Drupa, PHP, RDF, Python…	Non connu	Github / Gitlab/ Nakala et diffusion sous licence ouverte – Voir le PGD particulier du livrable	Voir le PGD particulier du livrable	Voir le PGD particulier du livrable – Actualisation du modèle de données du format pivot XML si la base cible est intégrée au portail

Produit de recherche	Description	Équipe	Cluster (volet B)	Nature des données	Formats / standards	Volumétrie	Politique de partage	Politique de conservation à long terme	Actions de Fairisation à mener
Développements réalisés par les équipes partenaires et intégrés à des systèmes d'information existant ou configuration particulières d'outils	Plateforme GED (CC)	voir ci-contre	-	codes informatiques	Javascript, IIIF, Drupa, PHP, RDF, Python…	Non connu	Github / Gitlab/ Nakala et diffusion sous licence ouverte – Voir le PGD particulier du livrable	Voir le PGD particulier du livrable	Voir le PGD particulier du livrable – Actualisation du modèle de données du format pivot XML si la base cible est intégrée au portail
Développements réalisés par les équipes partenaires et intégrés à des systèmes d'information existant ou configuration particulières d'outils	Implémentation des API IIIF dans les bibliothèques numériques des partenaires	voir ci-contre	-	codes informatiques	Javascript, IIIF, Drupa, PHP, RDF, Python…	Non connu	Github / Gitlab/ Nakala et diffusion sous licence ouverte – Voir le PGD particulier du livrable	Voir le PGD particulier du livrable	Voir le PGD particulier du livrable – Actualisation du modèle de données du format pivot XML si la base cible est intégrée au portail

Produit de recherche	Description	Équipe	Cluster (volet B)	Nature des données	Formats / standards	Volumétrie	Politique de partage	Politique de conservation à long terme	Actions de Fairisation à mener
Développements réalisés par les équipes partenaires et intégrés à des systèmes d'information existant ou configuration particulières d'outils	Interfaces ISMI (IRHT)	voir ci-contre	-	codes informatiques	Javascript, IIIF, Drupa, PHP, RDF, Python…	Non connu	Github / Gitlab/ Nakala et diffusion sous licence ouverte – Voir le PGD particulier du livrable	Voir le PGD particulier du livrable	Voir le PGD particulier du livrable
Développements réalisés par les équipes partenaires et intégrés à des systèmes d'information existant ou configuration particulières d'outils	Alignements automatisés des bases utilisant les référentiels de l'orient chrétien (IRHT)	voir ci-contre	-	codes informatiques	Javascript, IIIF, Drupa, PHP, RDF, Python…	Non connu	Github / Gitlab/ Nakala et diffusion sous licence ouverte – Voir le PGD particulier du livrable	Voir le PGD particulier du livrable	Voir le PGD particulier du livrable

Produit de recherche	Description	Équipe	Cluster (volet B)	Nature des données	Formats / standards	Volumétrie	Politique de partage	Politique de conservation à long terme	Actions de Fairisation à mener
Développements réalisés par les équipes partenaires et intégrés à des systèmes d'information existant ou configuration particulières d'outils	Mécanismes d'automatisation des mises à jour de données Biblindex sur le portail Biblissima (HiSoMA)	voir ci-contre	-	codes informatiques	Javascript, IIIF, Drupa, PHP, RDF, Python…	Non connu	Github / Gitlab/ Nakala et diffusion sous licence ouverte – Voir le PGD particulier du livrable	Voir le PGD particulier du livrable	Voir le PGD particulier du livrable
Développements réalisés par les équipes partenaires et intégrés à des systèmes d'information existant ou configuration particulières d'outils	Développement de l'interopérabilité des bases de données du CRH	voir ci-contre	-	codes informatiques	Javascript, IIIF, Drupa, PHP, RDF, Python…	Non connu	Github / Gitlab/ Nakala et diffusion sous licence ouverte – Voir le PGD particulier du livrable	Voir le PGD particulier du livrable	Voir le PGD particulier du livrable

Produit de recherche	Description	Équipe	Cluster (volet B)	Nature des données	Formats / standards	Volumétrie	Politique de partage	Politique de conservation à long terme	Actions de Fairisation à mener
Développements réalisés par les équipes partenaires et intégrés à des systèmes d'information existant ou configuration particulières d'outils	Valorisation et mise en ligne des objets numériques (AOROC)	voir ci-contre	-	codes informatiques	Javascript, IIIF, Drupa, PHP, RDF, Python…	Non connu	Github / Gitlab/ Nakala et diffusion sous licence ouverte – Voir le PGD particulier du livrable	Voir le PGD particulier du livrable	Voir le PGD particulier du livrable
Développements réalisés par les équipes partenaires et intégrés à des systèmes d'information existant ou configuration particulières d'outils	Outils Sigiscript de reconnaissance automatique embarqué pour les sceaux (SAPRAT)	voir ci-contre	-	codes informatiques	Javascript, IIIF, Drupa, PHP, RDF, Python…	Non connu	Github / Gitlab/ Nakala et diffusion sous licence ouverte – Voir le PGD particulier du livrable	Voir le PGD particulier du livrable	Voir le PGD particulier du livrable

Produit de recherche	Description	Équipe	Cluster (volet B)	Nature des données	Formats / standards	Volumétrie	Politique de partage	Politique de conservation à long terme	Actions de Fairisation à mener
Développements réalisés par les équipes partenaires et intégrés à des systèmes d'information existant ou configuration particulières d'outils	API référentiel de noms de lieux (CJM)	voir ci-contre	-	codes informatiques	Javascript, IIIF, Drupa, PHP, RDF, Python…	Non connu	Github / Gitlab/ Nakala et diffusion sous licence ouverte – Voir le PGD particulier du livrable	Voir le PGD particulier du livrable	Voir le PGD particulier du livrable
Développements réalisés par les équipes partenaires et intégrés à des systèmes d'information existant ou configuration particulières d'outils	Sparql Endpoint du SIAF	voir ci-contre	-	codes informatiques	Javascript, IIIF, Drupa, PHP, RDF, Python…	Non connu	Github / Gitlab/ Nakala et diffusion sous licence ouverte – Voir le PGD particulier du livrable	Voir le PGD particulier du livrable	Voir le PGD particulier du livrable

Produit de recherche	Description	Équipe	Cluster (volet B)	Nature des données	Formats / standards	Volumétrie	Politique de partage	Politique de conservation à long terme	Actions de Fairisation à mener
Développements réalisés par les équipes partenaires et intégrés à des systèmes d'information existant ou configuration particulières d'outils	Développement d'outils de données d'analyse de matériaux (CRC)	voir ci-contre	C2	codes informatiques	Javascript, IIIF, Drupa, PHP, RDF, Python…	Non connu	Github / Gitlab/ Nakala et diffusion sous licence ouverte – Voir le PGD particulier du livrable	Voir le PGD particulier du livrable	Voir le PGD particulier du livrable
Développements réalisés par les équipes partenaires et intégrés à des systèmes d'information existant ou configuration particulières d'outils	Développement de l'outil d'extraction d'éléments de décor Extractor (IRHT)	voir ci-contre	C3	codes informatiques	Javascript, IIIF, Drupa, PHP, RDF, Python…	Non connu	Github / Gitlab/ Nakala et diffusion sous licence ouverte – Voir le PGD particulier du livrable	Voir le PGD particulier du livrable	Voir le PGD particulier du livrable

Produit de recherche	Description	Équipe	Cluster (volet B)	Nature des données	Formats / standards	Volumétrie	Politique de partage	Politique de conservation à long terme	Actions de Fairisation à mener
Développements réalisés par les équipes partenaires et intégrés à des systèmes d'information existant ou configuration particulières d'outils	Développement de la reconnaissance automatique des formes et des fontes pour l'identification des imprimeurs base BaTyr (CESR)	voir ci-contre	C3	codes informatiques	Javascript, IIIF, Drupa, PHP, RDF, Python…	Non connu	Github / Gitlab/ Nakala et diffusion sous licence ouverte – Voir le PGD particulier du livrable	Voir le PGD particulier du livrable	Voir le PGD particulier du livrable
Développements réalisés par les équipes partenaires et intégrés à des systèmes d'information existant ou configuration particulières d'outils	Développement d'eScriptorium et intégration de Kraken (AOROC)	voir ci-contre	C4	codes informatiques	Javascript, IIIF, Drupa, PHP, RDF, Python…	Non connu	Github / Gitlab/ Nakala et diffusion sous licence ouverte – Voir le PGD particulier du livrable	Voir le PGD particulier du livrable	Voir le PGD particulier du livrable

Produit de recherche	Description	Équipe	Cluster (volet B)	Nature des données	Formats / standards	Volumétrie	Politique de partage	Politique de conservation à long terme	Actions de Fairisation à mener
Développements réalisés par les équipes partenaires et intégrés à des systèmes d'information existant ou configuration particulières d'outils	Re-développement et intégration d'Archetype au sein d'eScriptorium (AOROC)	voir ci-contre	C4	codes informatiques	Javascript, IIIF, Drupa, PHP, RDF, Python…	Non connu	Github / Gitlab/ Nakala et diffusion sous licence ouverte – Voir le PGD particulier du livrable	Voir le PGD particulier du livrable	Voir le PGD particulier du livrable
Développements réalisés par les équipes partenaires et intégrés à des systèmes d'information existant ou configuration particulières d'outils	Interfaces pour les écritures de haut en bas (AOROC)	voir ci-contre	C4	codes informatiques	Javascript, IIIF, Drupa, PHP, RDF, Python…	Non connu	Github / Gitlab/ Nakala et diffusion sous licence ouverte – Voir le PGD particulier du livrable	Voir le PGD particulier du livrable	Voir le PGD particulier du livrable

Produit de recherche	Description	Équipe	Cluster (volet B)	Nature des données	Formats / standards	Volumétrie	Politique de partage	Politique de conservation à long terme	Actions de Fairisation à mener
Développements réalisés par les équipes partenaires et intégrés à des systèmes d'information existant ou configuration particulières d'outils	Développement de Multipal pour les systèmes graphiques non encore traités	voir ci-contre	C4	codes informatiques	Javascript, IIIF, Drupa, PHP, RDF, Python…	Non connu	Github / Gitlab/ Nakala et diffusion sous licence ouverte – Voir le PGD particulier du livrable	Voir le PGD particulier du livrable	Voir le PGD particulier du livrable
Développements réalisés par les équipes partenaires et intégrés à des systèmes d'information existant ou configuration particulières d'outils	Publication web TEI de Carmina Latina Epigraphica Moderna CLEM (CJM)	voir ci-contre	C5a	codes informatiques	Javascript, IIIF, Drupa, PHP, RDF, Python…	Non connu	Github / Gitlab/ Nakala et diffusion sous licence ouverte – Voir le PGD particulier du livrable	Voir le PGD particulier du livrable	Voir le PGD particulier du livrable

Produit de recherche	Description	Équipe	Cluster (volet B)	Nature des données	Formats / standards	Volumétrie	Politique de partage	Politique de conservation à long terme	Actions de Fairisation à mener
Développements réalisés par les équipes partenaires et intégrés à des systèmes d'information existant ou configuration particulières d'outils	Développement d'une plateforme collaborative d'enrichissement des données de Biblindex (HiSoMA)	voir ci-contre	C5b	codes informatiques	Javascript, IIIF, Drupa, PHP, RDF, Python…	Non connu	Github / Gitlab/ Nakala et diffusion sous licence ouverte – Voir le PGD particulier du livrable	Voir le PGD particulier du livrable	Voir le PGD particulier du livrable
Développements réalisés par les équipes partenaires et intégrés à des systèmes d'information existant ou configuration particulières d'outils	Développement d'interfaces de visualisation spatio-temporelle des données statistiques de Biblindex (HiSoMA)	voir ci-contre	C5b	codes informatiques	Javascript, IIIF, Drupa, PHP, RDF, Python…	Non connu	Github / Gitlab/ Nakala et diffusion sous licence ouverte – Voir le PGD particulier du livrable	Voir le PGD particulier du livrable	Voir le PGD particulier du livrable

Produit de recherche	Description	Équipe	Cluster (volet B)	Nature des données	Formats / standards	Volumétrie	Politique de partage	Politique de conservation à long terme	Actions de Fairisation à mener
Développements réalisés par les équipes partenaires et intégrés à des systèmes d'information existant ou configuration particulières d'outils	Outil de collation assistée + API et interface (CJM)	voir ci-contre	C5b	codes informatiques	Javascript, IIIF, Drupa, PHP, RDF, Python…	Non connu	Github / Gitlab/ Nakala et diffusion sous licence ouverte – Voir le PGD particulier du livrable	Voir le PGD particulier du livrable	Voir le PGD particulier du livrable
Développements réalisés par les équipes partenaires et intégrés à des systèmes d'information existant ou configuration particulières d'outils	Conception d'environnements adaptés aux différents types de sources anciennes, médiévales, Renaissance (littéraires, encyclopédiques, diplomati	voir ci-contre	C5b	codes informatiques	Javascript, IIIF, Drupa, PHP, RDF, Python…	Non connu	Github / Gitlab/ Nakala et diffusion sous licence ouverte – Voir le PGD particulier du livrable	Voir le PGD particulier du livrable	Voir le PGD particulier du livrable

Produit de recherche	Description	Équipe	Cluster (volet B)	Nature des données	Formats / standards	Volumétrie	Politique de partage	Politique de conservation à long terme	Actions de Fairisation à mener
	ques, sources de la pratique), éditées par le CRAHAM et outillage de ces sources								
Développements réalisés par les équipes partenaires et intégrés à des systèmes d'information existant ou configuration particulières d'outils	Intégration de Collatinus et Eulexis dans la chaîne de traitement des données textuelles de Biblindex (HiSoMA)	voir ci-contre	C7	codes informatiques	Javascript, IIIF, Drupa, PHP, RDF, Python…	Non connu	Github / Gitlab/ Nakala et diffusion sous licence ouverte – Voir le PGD particulier du livrable	Voir le PGD particulier du livrable	Voir le PGD particulier du livrable
Développements réalisés par les équipes partenaires et intégrés à	Développement d'un outil générique de repérage de l'intertext	voir ci-contre	C7	codes informatiques	Javascript, IIIF, Drupa, PHP, RDF, Python…	Non connu	Github / Gitlab/ Nakala et diffusion sous licence ouverte – Voir le	Voir le PGD particulier du livrable	Voir le PGD particulier du livrable

Produit de recherche	Description	Équipe	Cluster (volet B)	Nature des données	Formats / standards	Volumétrie	Politique de partage	Politique de conservation à long terme	Actions de Fairisation à mener
des systèmes d'information existant ou configuration particulières d'outils	ualité (HiSoMA)						PGD particulier du livrable		
Développements réalisés par les équipes partenaires et intégrés à des systèmes d'information existant ou configuration particulières d'outils	Exports statiques ou dynamiques pour les échanges de données avec les partenaires et grandes infrastructures nationales	Persée, CCfr (Bnf), Biblindex (HiSoMA), OpenEdition, projet Triple (OPERAs), Huma-Num, GED (CC), Sudoc (via GED), FNE (ABES), ISIN, Geonames, etc. via projet CPER Condorcum	-	données textuelles formalisées	XML, CSV, varia	non connu	N / A	Voir le PGD particulier du livrable	Voir le PGD particulier du livrable

Produit de recherche	Description	Équipe	Cluster (volet B)	Nature des données	Formats / standards	Volumétrie	Politique de partage	Politique de conservation à long terme	Actions de Fairisation à mener
Enrichissement d'éditions, de corpus ou de bases de données des partenaires : Nouvelles notices, enregistrements, transcriptions, métadonnées, images, etc.	Bases héraldique et sigillographie (SAPRAT)	voir ci-contre	-	données textuelles, données textuelles formalisées, images 2D, images 3D	XML, CSV, varia	non connu	Voir le PGD particulier du livrable	Voir le PGD particulier du livrable	Voir le PGD particulier du livrable
Enrichissement d'éditions, de corpus ou de bases de données des partenaires : Nouvelles notices, enregistrements, transcriptions, métadon	Publication des nouvelles notices de reliures médiévales CRMBF dans la base Bibale (IRHT)	voir ci-contre	-	données textuelles, données textuelles formalisées, images 2D, images 3D	XML, CSV, varia	non connu	Voir le PGD particulier du livrable	Voir le PGD particulier du livrable	Voir le PGD particulier du livrable

Produit de recherche	Description	Équipe	Cluster (volet B)	Nature des données	Formats / standards	Volumétrie	Politique de partage	Politique de conservation à long terme	Actions de Fairisation à mener
nées, images, etc.									
Enrichissement d'éditions, de corpus ou de bases de données des partenaires : Nouvelles notices, enregistrements, transcriptions, métadonnées, images, etc.	Alimentation de la base filigranes (CJM)	voir ci-contre	-	données textuelles, données textuelles formalisées, images 2D, images 3D	XML, CSV, varia	non connu	Voir le PGD particulier du livrable	Voir le PGD particulier du livrable	Voir le PGD particulier du livrable
Enrichissement d'éditions, de corpus ou de bases de données des partenaires : Nouvelles notices, enregistre	Référencement et indexation d'empreintes de monnaies présentées sur des supports céramiques (CRAHAM)	voir ci-contre	-	données textuelles, données textuelles formalisées, images 2D, images 3D	XML, CSV, varia	non connu	Voir le PGD particulier du livrable	Voir le PGD particulier du livrable	Voir le PGD particulier du livrable

Produit de recherche	Description	Équipe	Cluster (volet B)	Nature des données	Formats / standards	Volumétrie	Politique de partage	Politique de conservation à long terme	Actions de Fairisation à mener
ments, transcriptions, métadonnées, images, etc.									
Enrichissement d'éditions, de corpus ou de bases de données des partenaires : Nouvelles notices, enregistrements, transcriptions, métadonnées, images, etc.	Acquisition des données numismatiques en musée et réserves et mise en ligne (AOROC)	voir ci-contre	-	données textuelles, données textuelles formalisées, images 2D, images 3D	XML, CSV, varia	non connu	Voir le PGD particulier du livrable	Voir le PGD particulier du livrable	Voir le PGD particulier du livrable
Enrichissement d'éditions, de corpus ou de bases de données des partenair	Intégration de données épigraphiques sur la base en ligne (AOROC)	voir ci-contre	-	données textuelles, données textuelles formalisées, images 2D, images 3D	XML, CSV, varia	non connu	Voir le PGD particulier du livrable	Voir le PGD particulier du livrable	Voir le PGD particulier du livrable

Produit de recherche	Description	Équipe	Cluster (volet B)	Nature des données	Formats / standards	Volumétrie	Politique de partage	Politique de conservation à long terme	Actions de Fairisation à mener
es : Nouvelles notices, enregistrements, transcriptions, métadonnées, images, etc.									
Enrichissement d'éditions, de corpus ou de bases de données des partenaires : Nouvelles notices, enregistrements, transcriptions, métadonnées, images, etc.	Géoréférencement et spatialisation des données épigraphiques sur Chronocarto (AOROC)	voir ci-contre	-	données textuelles, données textuelles formalisées, images 2D, images 3D	XML, CSV, varia	non connu	Voir le PGD particulier du livrable	Voir le PGD particulier du livrable	Voir le PGD particulier du livrable
Enrichissement d'éditions, de corpus ou de	Enrichissement des référentiels d'auteurs	voir ci-contre	-	données textuelles, données textuelles formalisées,	XML, CSV, varia	non connu	Voir le PGD particulier du livrable	Voir le PGD particulier du livrable	Voir le PGD particulier du livrable

Produit de recherche	Description	Équipe	Cluster (volet B)	Nature des données	Formats / standards	Volumétrie	Politique de partage	Politique de conservation à long terme	Actions de Fairisation à mener
bases de données des partenaires : Nouvelles notices, enregistrements, transcriptions, métadonnées, images, etc.	, oeuvres, noms de personnes, noms de lieux, matières avec des traits liés à la numismatique (CRAHAM, PDN de la MRSH de Caen)			images 2D, images 3D					
Enrichissement d'éditions, de corpus ou de bases de données des partenaires : Nouvelles notices, enregistrements, transcriptions, métadonnées, images, etc.	Expertises typologiques diverses sur les textes latins antiques et médiévaux. Production de données et de documentation érudite sur ces textes, dont édition,	voir ci-contre	-	données textuelles, données textuelles formalisées, images 2D, images 3D	XML, CSV, varia	non connu	Voir le PGD particulier du livrable	Voir le PGD particulier du livrable	Voir le PGD particulier du livrable

Produit de recherche	Description	Équipe	Cluster (volet B)	Nature des données	Formats / standards	Volumétrie	Politique de partage	Politique de conservation à long terme	Actions de Fairisation à mener
	transcription et critique de textes. Référentiels d'autorités pour le monde latin. (IRHT)								
Enrichissement d'éditions, de corpus ou de bases de données des partenaires : Nouvelles notices, enregistrements, transcriptions, métadonnées, images, etc.	Chaînage des outils d'édition et d'étude des documents d'archives (cartulaires et chartriers) : acquisition des données textuelles (CJM)	voir ci-contre	-	données textuelles, données textuelles formalisées, images 2D, images 3D	XML, CSV, varia	non connu	Voir le PGD particulier du livrable	Voir le PGD particulier du livrable	Voir le PGD particulier du livrable
Enrichissement d'éditions, de corpus ou de	Acquisition de transcriptions pour Epistémon (CESR)	voir ci-contre	-	données textuelles, données textuelles formalisées,	XML, CSV, varia	non connu	Voir le PGD particulier du livrable	Voir le PGD particulier du livrable	Voir le PGD particulier du livrable

Produit de recherche	Description	Équipe	Cluster (volet B)	Nature des données	Formats / standards	Volumétrie	Politique de partage	Politique de conservation à long terme	Actions de Fairisation à mener
bases de données des partenaires : Nouvelles notices, enregistrements, transcriptions, métadonnées, images, etc.				images 2D, images 3D					
Enrichissement d'éditions, de corpus ou de bases de données des partenaires : Nouvelles notices, enregistrements, transcriptions, métadonnées, images, etc.	Encodage TEI - MEI (CESR)	voir ci-contre	-	données textuelles, données textuelles formalisées, images 2D, images 3D	XML, CSV, varia	non connu	Voir le PGD particulier du livrable	Voir le PGD particulier du livrable	Voir le PGD particulier du livrable
Enrichissement	Edition et annotatio	voir ci-contre	-	données textuelles	XML, CSV,	non connu	Voir le PGD	Voir le PGD	Voir le PGD

Produit de recherche	Description	Équipe	Cluster (volet B)	Nature des données	Formats / standards	Volumétrie	Politique de partage	Politique de conservation à long terme	Actions de Fairisation à mener
d'éditions, de corpus ou de bases de données des partenaires : Nouvelles notices, enregistrements, transcriptions, métadonnées, images, etc.	n de sources (MRSH de Caen)			, données textuelles formalisées, images 2D, images 3D	varia		particulier du livrable	particulier du livrable	particulier du livrable
Enrichissement d'éditions, de corpus ou de bases de données des partenaires : Nouvelles notices, enregistrements, transcriptions, métadonnées,	Cartographie du patrimoine musical : enrichissements du fonds documentaire, étude des sources, exploitations du corpus (CESR)	voir ci-contre	-	données textuelles, données textuelles formalisées, images 2D, images 3D	XML, CSV, varia	non connu	Voir le PGD particulier du livrable	Voir le PGD particulier du livrable	Voir le PGD particulier du livrable

Produit de recherche	Description	Équipe	Cluster (volet B)	Nature des données	Formats / standards	Volumétrie	Politique de partage	Politique de conservation à long terme	Actions de Fairisation à mener
images, etc.									
Enrichissement d'éditions, de corpus ou de bases de données des partenaires : Nouvelles notices, enregistrements, transcriptions, métadonnées, images, etc.	Annotation et fouille des données musicales (CESR)	voir ci-contre	-	données textuelles, données textuelles formalisées, images 2D, images 3D	XML, CSV, varia	non connu	Voir le PGD particulier du livrable	Voir le PGD particulier du livrable	Voir le PGD particulier du livrable
Enrichissement d'éditions, de corpus ou de bases de données des partenaires : Nouvelles notices, enregistrements,	Préparation de corpus de textes patristiques (grec, latin, syriaque) pour utilisation du protocole DTS	voir ci-contre	-	données textuelles, données textuelles formalisées, images 2D, images 3D	XML, CSV, varia	non connu	Voir le PGD particulier du livrable	Voir le PGD particulier du livrable	Voir le PGD particulier du livrable

Produit de recherche	Description	Équipe	Cluster (volet B)	Nature des données	Formats / standards	Volumétrie	Politique de partage	Politique de conservation à long terme	Actions de Fairisation à mener
transcriptions, métadonnées, images, etc.									
Enrichissement d'éditions, de corpus ou de bases de données des partenaires : Nouvelles notices, enregistrements, transcriptions, métadonnées, images, etc.	Mise en oeuvre d'outils stylométriques et textométriques de repérage d'intertextualité sur des textes latins médiévaux (HiSoMA)	voir ci-contre	-	données textuelles, données textuelles formalisées, images 2D, images 3D	XML, CSV, varia	non connu	Voir le PGD particulier du livrable	Voir le PGD particulier du livrable	Voir le PGD particulier du livrable
Enrichissement d'éditions, de corpus ou de bases de données des partenaires :	Numérisation et océrisation de textes latins sur imprimés anciens (IRHT)	voir ci-contre	-	Images, données textuelles (OCR brut et corrigé) – données textuelles, données numériques,	format texte, XML, modèles 3D	non connu	Voir le PGD particulier du livrable - souvent : Nakala	Voir le PGD particulier du livrable	Voir le PGD particulier du livrable

Produit de recherche	Description	Équipe	Cluster (volet B)	Nature des données	Formats / standards	Volumétrie	Politique de partage	Politique de conservation à long terme	Actions de Fairisation à mener
Nouvelles notices, enregistrements, transcriptions, métadonnées, images, etc.				données image 2D et 3D					
Enrichissement d'éditions, de corpus ou de bases de données des partenaires : Nouvelles notices, enregistrements, transcriptions, métadonnées, images, etc.	Numérisation et océrisation de documents patrimoniaux et d'archives de la recherche (GED)	voir ci-contre	-	Images, données textuelles (OCR brut et corrigé) – données textuelles, données numériques, données image 2D et 3D	format texte, XML, modèles 3D	non connu	Voir le PGD particulier du livrable - souvent : Nakala	Voir le PGD particulier du livrable	Voir le PGD particulier du livrable
Enrichissement d'éditions, de corpus ou de bases de	Numérisations ponctuelles d'échantillon d'édition	voir ci-contre	-	Images, données textuelles (OCR brut et corrigé) – données	format texte, XML, modèles 3D	non connu	Voir le PGD particulier du livrable - souvent : Nakala	Voir le PGD particulier du livrable	Voir le PGD particulier du livrable

Produit de recherche	Description	Équipe	Cluster (volet B)	Nature des données	Formats / standards	Volumétrie	Politique de partage	Politique de conservation à long terme	Actions de Fairisation à mener
données des partenaires : Nouvelles notices, enregistrements, transcriptions, métadonnées, images, etc.	s et archives de chercheurs (CESR)			textuelles, données numériques, données image 2D et 3D					
Enrichissement d'éditions, de corpus ou de bases de données des partenaires : Nouvelles notices, enregistrements, transcriptions, métadonnées, images, etc.	Numérisations 3D d'inscriptions médiévales (CESCM)	voir ci-contre	-	Images, données textuelles (OCR brut et corrigé) – données textuelles, données numériques, données image 2D et 3D	format texte, XML, modèles 3D	non connu	Voir le PGD particulier du livrable - souvent : Nakala	Voir le PGD particulier du livrable	Voir le PGD particulier du livrable
Enrichissement d'édition	Numérisations 3D de	voir ci-contre	-	Images, données textuelles	format texte, XML,	non connu	Voir le PGD particulie	Voir le PGD particulie	Voir le PGD particulie

Produit de recherche	Description	Équipe	Cluster (volet B)	Nature des données	Formats / standards	Volumétrie	Politique de partage	Politique de conservation à long terme	Actions de Fairisation à mener
s, de corpus ou de bases de données des partenaires : Nouvelles notices, enregistrements, transcriptions, métadonnées, images, etc.	reliures médiévales - projet CRMBF-3D (IRHT)			(OCR brut et corrigé) – données textuelles, données numériques, données image 2D et 3D	modèles 3D		r du livrable - souvent : Nakala	r du livrable	r du livrable
Enrichissement d'éditions, de corpus ou de bases de données des partenaires : Nouvelles notices, enregistrements, transcriptions, métadonnées, images, etc.	Campagne photographique pour le répertoire de filigranes (IRHT)	voir ci-contre	-	Images, données textuelles (OCR brut et corrigé) – données textuelles, données numériques, données image 2D et 3D	format texte, XML, modèles 3D	non connu	Voir le PGD particulier du livrable - souvent : Nakala	Voir le PGD particulier du livrable	Voir le PGD particulier du livrable

Produit de recherche	Description	Équipe	Cluster (volet B)	Nature des données	Formats / standards	Volumétrie	Politique de partage	Politique de conservation à long terme	Actions de Fairisation à mener
Enrichissement d'éditions, de corpus ou de bases de données des partenaires : Nouvelles notices, enregistrements, transcriptions, métadonnées, images, etc.	Numérisation des photographies / dessins d'inscriptions (AOROC)	voir ci-contre	-	Images, données textuelles (OCR brut et corrigé) – données textuelles, données numériques, données image 2D et 3D	format texte, XML, modèles 3D	non connu	Voir le PGD particulier du livrable - souvent : Nakala	Voir le PGD particulier du livrable	Voir le PGD particulier du livrable
Enrichissement d'éditions, de corpus ou de bases de données des partenaires : Nouvelles notices, enregistrements, transcriptions, métadon	Traitement numérique de la donnée musicale / Reconstitution d'espaces sonores (CESR)	voir ci-contre	-	Images, données textuelles (OCR brut et corrigé) – données textuelles, données numériques, données image 2D et 3D	format texte, XML, modèles 3D	non connu	Voir le PGD particulier du livrable - souvent : Nakala	Voir le PGD particulier du livrable	Voir le PGD particulier du livrable

Produit de recherche	Description	Équipe	Cluster (volet B)	Nature des données	Formats / standards	Volumétrie	Politique de partage	Politique de conservation à long terme	Actions de Fairisation à mener
nées, images, etc.									
Enrichissement d'éditions, de corpus ou de bases de données des partenaires : Nouvelles notices, enregistrements, transcriptions, métadonnées, images, etc.	Projet Relicantus : inventaire et numérisation de fragments musicaux – campagne de numérisation (IRHT)	voir ci-contre	-	Images, données textuelles (OCR brut et corrigé) – données textuelles, données numériques, données image 2D et 3D	format texte, XML, modèles 3D	non connu	Voir le PGD particulier du livrable - souvent : Nakala	Voir le PGD particulier du livrable	Voir le PGD particulier du livrable
Enrichissement d'éditions, de corpus ou de bases de données des partenaires : Nouvelles notices, enregistre	Analyses physico-chimiques des supports de l'écrit (sur papyrus, parchemin, papier, pierre, monnaies, objets	AOROC	-	données textuelles	non connu	non connu	Voir le PGD particulier du livrable	Voir le PGD particulier du livrable	Voir le PGD particulier du livrable

Produit de recherche	Description	Équipe	Cluster (volet B)	Nature des données	Formats / standards	Volumétrie	Politique de partage	Politique de conservation à long terme	Actions de Fairisation à mener
ments, transcriptions, métadonnées, images, etc.	métalliques)								
Enrichissement d'éditions, de corpus ou de bases de données des partenaires : Nouvelles notices, enregistrements, transcriptions, métadonnées, images, etc.	Supports de présentation, vidéos	Équipe portail	-	documents web et multimédia	.pdf, .ppt, .mp4	plus de 180 ressources	Publié en ligne sur le site projet.biblissima.fr (Biblissima 1) et sur Zenodo (Biblissima+)	N / A	Non
MOOCs et démonstrateurs	Cours d'auto-formation à l'encodage et à la publication de sources TEI(CIHA	voir ci-contre (Cluster 5b)	-	documents web et multimédia	à définir	4 cours, quelques ressources	Publié en ligne sous licence CC BY et accessible via le site Biblissima	N / A	Non

Produit de recherche	Description	Équipe	Cluster (volet B)	Nature des données	Formats / standards	Volumétrie	Politique de partage	Politique de conservation à long terme	Actions de Fairisation à mener
	M, HiSoMA), formations TEI débutant et avancé (CESR) Formations DTS (CJM)								

17 6 – Ressources du périmètre P3

Ce tableau recense les jeux de données produits par les projets lauréats de l'appel à manifestation d'intérêt annuel (2022-2027).

Type de données brutes fournies	Nom	Équipe fondatrice associée	Catégorie de soutien	Type de ressource du projet global selon catégories du Portail	Utilisation native des référentiels Biblissima	Suivi intégration	Licence de diffusion du partenaire pour la base source	Politique de conservation à long terme du partenaire pour la base source	Lien
XML-ALTO (données d'entrainement)	Fabliaux	CIHAM	Projet partenarial	Corpus ou édition de source	oui	N / A	CC-BY 4.0	non	https://github.com/CIHAM-HTR/Fabliaux
images .jpg	Fabliaux	CIHAM	Projet partenarial	Corpus ou édition de source	oui	N / A	CC-BY 4.0	non	https://github.com/CIHAM-HTR/Fabliaux
XML-TEI	Fabliaux	CIHAM	Projet partenarial	Corpus ou édition de source	oui, pour les métadonnées des manuscrits	B+, non intégré	Etalab + CC-BY 4.0	Dépôt sur Nakala prévu fin 2023	https://gitlab.huma-num.fr/fabliaux/public/fabliaux-textes-diffusion/-/tree/main/TEI
TXT	Fabliaux	CIHAM	Projet partenarial	Corpus ou édition de source	non	N / A	Etalab	non	https://gitlab.huma-num.fr/fabliaux/public/fabliaux-textes-diffusion/-/tree/main/TXT

Type de données brutes fournies	Nom	Équipe fondatrice associée	Catégorie de soutien	Type de ressource du projet global selon catégories du Portail	Utilisation native des référentiels Biblissima	Suivi intégration	Licence de diffusion du partenaire pour la base source	Politique de conservation à long terme du partenaire pour la base source	Lien
Mlmodel (model HTR Kraken)	Fabliaux	CIHAM	Projet partenarial	Corpus ou édition de source	non	N / A	CC-BY 4.0	non	https://gitlab.huma-num.fr/fabliaux/members/fabliaux_htr_models
XML-Alto (prédiction HTR)	Fabliaux	CIHAM	Projet partenarial	Corpus ou édition de source	non	N / A	CC-BY 4.0	non	https://gitlab.huma-num.fr/fabliaux/members/fabliaux_htr_prediction
non connu	Bibliotheca Carnotensis Nova	IRHT	Projet partenarial	Catalogue ou répertoire	non connu	non connu	non connu	non connu	-
non connu	Éditions critiques relatives à l'Université de Paris (ECRU)	IRHT	Projet partenarial	Corpus ou édition de source	non connu	non connu	non connu	non connu	-
non connu	Manuscrits	IRHT	Projet partenari	Corpus ou	non connu	non connu	non connu	non connu	-

Type de données brutes fournies	Nom	Équipe fondatrice associée	Catégorie de soutien	Type de ressource du projet global selon catégories du Portail	Utilisation native des référentiels Biblissima	Suivi intégration	Licence de diffusion du partenaire pour la base source	Politique de conservation à long terme du partenaire pour la base source	Lien
	Génovéfains (MAGE)		al	édition de source					
non connu	RESCAPÉ	CJM	Projet partenarial	Corpus ou édition de source	non connu	non connu	non connu	non connu	-
non connu	Reverse Engineering Kennicott (REK)	AOrOc	Projet partenarial	Base de données scientifiques	non connu	non connu	non connu	non connu	-
non connu	Smart Critical Ben Sira (SCRIBES)	CRAHAM	Projet partenarial	Corpus ou édition de source	non connu	non connu	non connu	non connu	-
non connu	Venezia Libro Aperto (VeLA)	CESCM	Projet partenarial	Corpus ou édition de source	non connu	non connu	non connu	non connu	-
non connu	Du papyrus hiératique au texte numérique	HiSoMA	Bourse Jeune Chercheur	Corpus ou édition de source	non connu	non connu	non connu	non connu	-
non connu	Un catalogue	HiSoMA	Bourse Jeune	Corpus ou	non connu	non connu	non connu	non connu	-

Type de données brutes fournies	Nom	Équipe fondatrice associée	Catégorie de soutien	Type de ressource du projet global selon catégories du Portail	Utilisation native des référentiels Biblissima	Suivi intégration	Licence de diffusion du partenaire pour la base source	Politique de conservation à long terme du partenaire pour la base source	Lien
	des travaux savants imprimés et manuscrits consacrés à Euripide au XVIe siècle		Chercheur	édition de source					
non connu	EpiHorrea	AOrOc	Projet exploratoire	Corpus ou édition de source	non connu	non connu	non connu	non connu	-
non connu	HTRogène	CIHAM	Projet exploratoire	Corpus ou édition de source	non connu	non connu	non connu	non connu	-
non connu	Manuscrits des classiques latins de la bibliothèque Farnèse	MRSH-PDN	Projet exploratoire	Catalogue ou répertoire	non connu	non connu	non connu	non connu	-
non connu	Burgundia Scripta	IRHT	Projet partenarial	Corpus ou	non connu	non connu	non connu	non connu	-

Type de données brutes fournies	Nom	Équipe fondatrice associée	Catégorie de soutien	Type de ressource du projet global selon catégories du Portail	Utilisation native des référentiels Biblissima	Suivi intégration	Licence de diffusion du partenaire pour la base source	Politique de conservation à long terme du partenaire pour la base source	Lien
	Merovingica			édition de source					
non connu	Hyper-Estampages	HiSoMA	Projet partenarial	Catalogue ou répertoire	non connu	non connu	non connu	non connu	-
non connu	Les dessins de sceaux de la collection Gaignières	IRHT	Projet partenarial	Catalogue ou répertoire	non connu	non connu	non connu	non connu	-
non connu	Les tablatures de la collection Albani	CESR	Projet partenarial	Base de données scientifiques	non connu	non connu	non connu	non connu	-
non connu	SION Digit Sfarim	IRHT	Projet partenarial	Base de données scientifiques	non connu	non connu	non connu	non connu	-

IV. PGD détaillé (Périmètre 1)

18 PGD détaillé de l'infrastructure numérique

Cette partie développe le PGD détaillé du périmètre de données principal de l'infrastructure numérique de **Biblissima+** (périmètre P1) selon le plan du modèle générique de PGD de l'**ANR**. Pour une vue synthétique sur les données produites ou collectées dans ce périmètre, se reporter aux tableaux 1 et 2 de la partie précédente « Vue d'ensemble des jeux de données ».

19 Description des données et collecte ou réutilisation de données existantes

19.1 A/ Recueil de nouvelles données et réutilisation de données existantes

Depuis son lancement en 2017 au cours de la première période de financement ÉquipEx, le portail Biblissima s'est enrichi par vagues successives d'intégration de nouveaux jeux de données. Ces jeux de données sont issus des catalogues, des bases de données scientifiques, des opérations de numérisation qui sont menées par les équipes partenaires ainsi que du moissonnage de collections d'images publiées sur le web à l'aide du standard **IIIF**. Ce fonctionnement est conservé dans le cadre de **Biblissima+**. Il relève à la fois de la collecte de données existantes et de la création de nouvelles données. En effet, à chaque intégration d'une ressource au sein du cluster de données sous-jacent au portail web, un processus de normalisation et d'enrichissement des données est mis en œuvre. Les entités déjà présentes dans les référentiels d'autorité sont indexées par les identifiants de *data.biblissima*, tandis que de nouveaux identifiants sont créés pour les entités inédites au sein du référentiel, ce qui l'enrichit en retour. Pour bien comprendre les étapes des traitements et la manière dont ils s'articulent entre eux, il est nécessaire de décrire la manière dont ces référentiels sont gérés, utilisés et enrichis à chaque campagne d'ingestion d'une ressource, de même que la chaîne opératoire dans sa globalité. Les mécanismes de mise à jour à définir dans le cadre de **Biblissima+** devront en effet s'appuyer sur cette chaîne opératoire et la consolider au niveau technique et méthodologique.

19.1.1 Les référentiels Biblissima, l'épine dorsale des mécanismes d'interopérabilité des données

Le traitement des différents types de données transmises par les partenaires du projet s'accompagne de la création de référentiels qui servent à gérer l'ensemble des données et facilitent l'intégration progressive des bases dans le portail Biblissima. Des référentiels ont ainsi été créés pour les personnes et les collectivités, pour les œuvres, les lieux géographiques, les descripteurs iconographiques, les cotes de manuscrits et imprimés. Ils contiennent des formes graphiques préférentielles et alternatives, l'indication de la langue d'origine dans le cas des œuvres, des notes d'identification en provenance des partenaires ou rédigées par l'équipe Biblissima, des liens vers les pages source des données si elles sont disponibles, des identifiants uniques de type **ARK**, et des alignements vers des jeux de données liées (*Linked Open Data* mis en libre accès par différentes institutions et projets – **BnF**, **Library of Congress**, **DNB**, **SUDOC**, **VIAF**, **Wikidata**, **Geonames**, **Pleiades**, **Trismegistos** etc.). Ces référentiels sont publiés sous licence ouverte via la plateforme *data.biblissima.fr*. Ils sont mis à disposition sous une forme structurée et exploitable par des programmes informatiques via des services web (ou API web). Tout projet intéressé a ainsi la possibilité de les récupérer et de les réutiliser en utilisant l'un des points d'accès proposés[1].

19.1.2 Harmonisation, alignements et enrichissements de données

La mise en interopérabilité des jeux de données hétérogènes au sein portail **Biblissima+** repose sur un processus de traitement en 4 grandes étapes : conversion (ou récupération) dans un format dit « pivot », extraction et normalisation des entités, alignements vers des ressources externes et enrichissement des données initiales en retour. Le but est d'agréger plusieurs types de données en provenance des partenaires du projet (périmètres P2 et P3). Au départ du processus, un jeu de données à intégrer est structuré selon différents modèles et formats (**SQL**, **MARC**, **TEI**, **EAD**…). Il est d'abord transformé vers un même modèle pivot au format XML qui a été défini par l'équipe technique du projet Biblissima. Une fois cette transformation opérée, les données (entités) de chaque base de données sont extraites, retraitées le cas échéant et alignées les unes avec les autres afin de regrouper dans un seul point d'accès les différentes formes graphiques d'une entité (autrement dit toutes les formes du nom d'une même personne, d'une même cote de manuscrit, d'une même œuvre, etc.). Il s'agit d'une étape d'harmonisation essentielle qui permet de regrouper le maximum d'informations pertinentes relatives à une même entité dans une grappe bien identifiée au sein du cluster de données. Des alignements vers des jeux de données liées (*Linked Open Data*) sont également ajoutés. Ils sont utilisés pour récupérer des informations structurées (éléments biographiques, formes graphiques alternatives, ou encore coordonnées géographiques) qui viennent compléter les données initialement reçues et contribuent à les rendre interopérables avec d'autres outils ou d'autres vocabulaires (cf. plus haut). Par exemple, c'est ce mécanisme qui permettra de lier les entités à des lieux géographiques représentables sur une carte. Cette étape d'alignement vers des sources externes et d'enrichissement des données permet aussi d'inscrire le portail Biblissima dans l'écosystème plus large du web sémantique.

19.1.3 Mises à jour du portail

Les jeux de données sont périodiquement récupérés des partenaires du consortium Biblissima et versés dans le portail dans des délais dépendants de la charge de travail de l'équipe portail. La volumétrie des bases et le temps de traitement afférent représentent des facteurs qui dépendent des spécificités de chaque base et dont le calendrier précis d'intégration dans le portail ne peut pas être défini a priori. Afin de faciliter l'étape de traitement, il est utile que les bases partenaires s'appuient sur les référentiels Biblissima. Au cas où les référentiels ne disposent pas d'une entité équivalente, les partenaires doivent la créer dans la plateforme *data.biblissima* avec l'aide de l'équipe portail. Cette étape peut être faite soit manuellement, soit par versement à partir d'un fichier tabulaire, soit automatiquement via l'API de la plateforme **Wikibase** gérant les référentiels. L'insertion des identifiants Biblissima dans chacune des bases partenaires facilite le processus de recoupement des informations, augmente l'interopérabilité des données dans le portail et améliore le rythme des mises à jour.

Dans le cadre de **Biblissima+**, de nouvelles procédures sont susceptibles de simplifier les mises à jour :

- Mettre en place un web service qui permette de récupérer les données à tout moment (en veillant à ce qu'il soit facilement aligné ou alignable avec le format pivot) ;
- Déposer à intervalles réguliers (par exemple tous les 4 ou 6 mois) les jeux de données sur une plateforme accessible à l'équipe portail (cf. plus haut sur l'utilisation de la plateforme Zenodo pour le stockage des données sources, en accès ouvert ou restreint).

19.1.4 Développements liés au portail

L'élargissement du périmètre de **Biblissima+** induit l'apparition de nouveaux types de données via les 7 clusters (données sur les matériaux, éditions **TEI**, transcriptions issues de l'**HTR** notamment). Leur prise en charge nécessitera des développements spécifiques du portail qui seront assurés partiellement en interne par l'équipe. Chaque nouveau type de données peut avoir des répercussions à plusieurs niveaux : sur le format pivot, les scripts de traitement, le modèle de données du portail et de *data.biblissima*, le module d'import des données dans le portail ou l'affichage de ces données dans les pages web du portail. Ces ajustements aux différentes étapes de la chaîne de traitement et de publication sont maîtrisés par l'équipe portail.

Un ensemble de développements liés à l'amélioration des fonctionnalités offertes par l'infrastructure ou à sa consolidation sont également prévus (moteur de recherche, facettes, visualisations de données, visualiseur d'images, exports à la demande, passerelles automatisées entre le portail et *data.biblissima* etc.). Ces évolutions fonctionnelles seront soit prises en charge par l'équipe dans la limite du temps et des compétences disponibles, soit feront l'objet de marchés de prestations informatiques.

19.2 B/ Description des données collectées et produites

Les données collectées sont transformées vers le format pivot XML qui a été modélisé à partir de modèles conceptuels et d'ontologies qui font référence pour le périmètre scientifique de Biblissima (**TEI**, **EAD**, **Cidoc-CRM**, **FRBR**). Pour le détail des natures, types, formats, standards et volumes de chaque source collectée ou produite et possiblement intégrée au cluster de données, se reporter aux tableaux de synthèse de la partie précédente.

---

1. 4 points d'accès sont proposés : une API Mediawiki/Wikibase, une interface de données liées (RDF), un point d'accès SPARQL (en test) et un service de réconciliation et d'alignement de données pour l'outil OpenRefine. ↩

20 Documentation et qualité des données

20.1 A/ Métadonnées et documentation accompagnant les données

Dans **Biblissima+**, les nouveaux jeux de données donneront lieu à des dépôts dans un entrepôt de données à différents stades de leur cycle de vie :

- Données « brutes » résultant d'un export, statique ou dynamique ;
- Données converties vers le format pivot, avec le fichier de mapping utilisé le cas échéant ;
- Données traitées et enrichies (après alignement et ajout d'informations).

Il est recommandé d'utiliser des entrepôts spécialisés dans le partage et l'archivage de données et attribuant des identifiant pérennes DOI tels que **Zenodo** – entrepôt du CERN financé par la Commission européenne – ou **Nakala** – réalisation de l'infrastructure de recherche **Huma-Num**. **Nakala** peut dans certain cas, et après un audit, donner accès à un archivage sur les serveurs du **CINES** en France. En ce qui concerne les codes sources de logiciels, le choix se porte sur l'archive pérenne de logiciels **Software Heritage**.

Chaque dépôt dans une plateforme de ce type donne lieu à une fiche de métadonnées structurée, conforme à un standard (Datacite pour la plateforme **Zenodo** – voir l'exemple donné en annexe – Dublin Core s'il s'agit de **Nakala** ; ou Codemeta pour **Software Heritage**).

Lors de la constitution des dépôts, la documentation nécessaire à l'intelligibilité des données est réunie ou produite. Il peut s'agir de fichiers texte de type README décrivant le dictionnaire des données, le modèle ou schéma utilisé, la licence d'utilisation, l'historique des traitements précédents ou toute information jugée utile pour comprendre l'organisation des fichiers dans le jeu. Il peut s'agir également de documents de spécifications, de schéma conceptuel de bases de données, de documentations de toutes sortes. La sélection des éléments de documentation pertinents ou la définition du niveau de détail apporté dans les métadonnées est à définir au cas par cas.

Il est à noter qu'une grande partie des travaux sont dès le départ fondés sur des standards : **TEI**, **EAD**, **IIIF**. Les données produites dans ce cadre sont nativement riches en métadonnées. Par exemple, la **TEI** et **EAD** comportent obligatoirement un en-tête de métadonnées descriptives structurées et riches en informations sur la provenance, la bibliographie, les choix d'encodage, les conventions éditoriales ou de transcription. Certains éléments comme des mots clés, des concepts, des descriptions peuvent être utilisés pour renseigner les métadonnées au niveau de la fiche de métadonnées du dépôt. Dans le même ordre d'idées, l'utilisation des référentiels d'autorité Biblissima nativement dans les catalogues, les corpus et les éditions renforcera la dimension *FAIR* des données et jeux de données.

Pour le détail des formats et standards utilisés par jeu de données, se reporter aux tableaux synthétiques de la partie précédente.

20.2 B/ Mesures de contrôle de la qualité des données

Le processus d'ingestion des ressources produites par les équipes partenaires dans le cluster de données du portail ayant pour but l'harmonisation, la normalisation et l'enrichissement des données s'appuient sur des scripts successifs et une vérification humaine. Il garantit ainsi un très haut niveau de qualité technique, syntaxique et sémantique des jeux de données mis en interopérabilité au sein du portail. La qualité scientifique des jeux de données est quant à elle vérifiée en amont, avant ingestion. L'information sur les processus qualité mis en œuvre par les équipes scientifiques est fournie dans les PGDs particuliers qui seront rédigés pour chaque livrable par les responsables scientifiques et techniques. Pour une vue d'ensemble des ressources produites dans les périmètres P2 et P3, se reporter aux tableaux synthétiques de la partie précédente.

21 Stockage et sauvegarde pendant le processus de recherche

21.1 A/ Stockage et politique de sauvegarde

Les données du périmètre P1 sont sauvegardées pendant le projet à différents niveaux[1].

21.1.1 Cluster de données

Les serveurs qui hébergent l'infrastructure du portail et ses différentes applications font l'objet de *snapshots* (instantanés) automatiques journaliers.

Les bases de données sous-jacentes du portail (PostgreSQL) et de *IIIF-Collections* (ElasticSearch) ne sont pas sauvegardées automatiquement, les *snapshots* évoqués ci-avant étant jugés suffisants. De toute façon ces bases de données restent « statiques », en ce sens qu'elles n'évoluent que lorsque l'on décide d'une mise à jour. Ainsi la dernière version des données importées dans ces bases est toujours ce qui fait foi, et peut à tout moment être réimportée. Ces données (XML pour le portail, et CSV pour *IIIF-Collections*) sont stockées dans l'espace *Seafile* de l'équipe hébergée au **Campus Condorcet**.

21.1.2 Méthodes, protocoles et scripts utilisés pour la normalisation, l'alignement, l'ingestion des ressources dans le portail et l'enrichissement des référentiels

Les codes informatiques de logiciels et scripts sont gérés via la plateforme d'hébergement de code source **Gitlab** (sur l'instance proposée par **Huma-Num**). Une partie des dépôts est gérée et diffusée sur la plateforme **Github** afin d'augmenter leur visibilité et faciliter d'éventuelles contributions extérieures.

Les dépôts de codes et l'ensemble des répertoires de travail sont présents sur les ordinateurs professionnels des membres de l'équipe portail, qui sont au nombre de 3 : - MacBook Pro (chef de projet), - MacBook Air (responsable des référentiels d'autorité et des traitements), - MacBook Pro (développeur).

Chaque membre de l'équipe dispose d'un disque dur externe pour la sauvegarde de ses fichiers, mais en principe l'intégralité des fichiers est toujours stockée dans un espace partagé sur les serveurs de l'infrastructure numérique du **Campus Condorcet** au sein du service de Drive *Seafile*. Tous ces fichiers sont synchronisés et donc sauvegardés à distance chaque jour. La politique de sauvegarde du **Campus Condorcet** en ce qui concerne *Seafile* n'est pas connue.

21.2 B/ Mesures concernant la sécurité des données et la protection des données sensibles

21.2.1 Récupération des données en cas d'incident

En cas d'incident, les données seront récupérées avec l'aide du Pôle numérique du **Campus Condorcet** en charge de l'administration système, en s'appuyant sur les *snapshots* effectués chaque jour pour chacun des serveurs de **Biblissima+**.

Tous les codes sources hébergés dans **Gitlab** ou **Github** seront récupérés à partir des dépôts distants et réinstallés facilement selon les procédures communes de ce type d'outils de gestion de codes informatiques.

21.2.2 Sécurité et protection de données sensibles

Le cluster de données ne comporte pas de données sensibles : les informations collectées ou produites ne révèlent pas la prétendue origine raciale ou ethnique, les opinions politiques, les convictions religieuses ou philosophiques ou l'appartenance syndicale. Elles ne comportent pas de données génétiques, de données biométriques, de données concernant la santé, la vie sexuelle ou l'orientation sexuelle d'une personne physique.

21.2.3 Droits d'accès

Le tableau ci-dessous détaille l'organisation des répertoires où sont stockées les données sur les serveurs du **Campus Condorcet** (Drive *Seafile*).

Type	Bibliothèque ou répertoire Seafile	Droits d’accès et niveau de droits
Cluster de données	bbma-data	Équipe technique Biblissima+
Codes informatiques	BB - Scripts	Équipe technique Biblissima+
Plateforme de référentiels	BB - Scripts/data.biblissima	Équipe technique Biblissima+
IIIF-Collections	BB - Scripts/iiif-index-data	Équipe technique Biblissima+
Espaces de travail de l’équipe portail	BB - Espace de travail	Équipe technique Biblissima+

1. L'hébergement de l’infrastructure technique est opéré par le Pôle Numérique de l'établissement public Campus Condorcet. ↩

22 Exigences légales et éthiques, codes de conduite

22.1 A/ Données à caractère personnel

Le référentiel d'autorité des personnes physiques de **Biblissima+** porte pour l'essentiel sur des individus décédés depuis plusieurs siècles. Le règlement européen sur la protection des données (**RGPD**) ne s'applique pas aux personnes décédées.

22.2 B/ Autres questions juridiques

Les questions de propriété juridique et de cession de droits d'exploitation ou de représentation sont traitées dans l'accord de consortium **Biblissima+** signé le 19 avril 2023.

22.2.1 Cluster de données du portail Biblissima+

Les nouvelles données intégrées au portail pendant le projet **Biblissima+** comporteront obligatoirement une licence explicite, exprimée dans un fichier texte de licence, facilement identifiable dans les dépôts qui seront exploités par l'équipe portail (voir plus haut). Cette manière de procéder garantit qu'aucun jeu de données n'est intégré sans licence explicite pour la diffusion et la réutilisation par le public via le portail **Biblissima+** et ses API.

Les données au format pivot et les données enrichies lors de la phase de traitement seront déposées sur la plateforme **Zenodo** en accès libre ou réservé, selon les conditions indiquées dans les fichiers de licence rédigés par les producteurs des données.

22.2.2 Moteur IIIF Collections

Les données exploitées dans le cadre du moteur *IIIF-Collections* respectent les conditions et droits d'utilisation explicitement indiqués par les collections sources. Les licences et autres informations d'attribution présentes dans les données à la source sont systématiquement affichées sur le site *IIIF-Collections*.

Pour ce qui concerne l'affichage des documents numérisés dans le visualiseur de **Biblissima+**, en l'absence de mentions de licence ou de restrictions particulières apposées par les institutions de conservation, on considère que la mise à disposition de leurs numérisations via les protocoles d'interopérabilité **IIIF** autorise de fait à les afficher dans le portail **Biblissima+,** étant donné que ces standards d'échange d'images sont spécialement conçus pour cela.

22.3 C/ Questions éthiques et codes déontologiques

Le cadre éthique et déontologique du projet **Biblissima+** dans le périmètre P1 porte essentiellement sur le respect de la propriété intellectuelle, définie dans la partie précédente sur les questions juridiques.

En ce qui concerne les jeux de données produits ou collectés dans le cadre des livrables, ces dimensions sont traitées dans les PGDs particuliers. Si une réflexion collective sur les sujets éthiques ou déontologiques est nécessaire à cause de la nature particulière des données, celle-ci pourra être menée dans le cadre des clusters thématiques.

23 Partage des données et conservation à long terme

23.1 A/ Périodes, modalités, restrictions ou embargos

Les modalités de partage et d'archivage varient selon la nature des données. Pour une vue d'ensemble de la politique de partage, se reporter aux tableaux de synthèse de la partie précédente.

Les jeux de données du périmètre 1 sont essentiellement constitués de codes sources informatiques ou de données textuelles formalisées (**TEI**, **XML**, **RDF**).

23.1.1 A.1 Codes sources

Les codes informatiques de logiciels et scripts sont gérés sur la plateforme Gitlab proposée par **Huma-Num**. D'autre part, une partie des dépôts est gérée et diffusée sur la plateforme **Github** afin d'augmenter leur visibilité et faciliter d'éventuelles contributions extérieures (code, remontée de bugs). Cela vaut surtout pour les logiciels ayant un fort potentiel de réutilisation en dehors du contexte de **Biblissima+**, ou pour lesquels des contributions de développeurs ou utilisateurs extérieurs sont souhaitées ou se sont déjà produites par le passé (c'est notamment le cas de Collatinus, qui dispose d'une petite communauté). Les dépôts issus de « forks » seront par essence eux aussi gérés sur **Github**.

Suivant le temps disponible, les besoins de citabilité du jeu ou l'impact recherché pour la réutilisation, 3 stratégies peuvent être adoptées :

Stratégie 1 : archivage « instantané » sans ajout de métadonnées, identifiant citable *SWHID*

- Les codes sources gérés sur une plateforme de gestion de code source (Gitlab de préférence, Github si le développement Open Source implique la communauté) peuvent être sauvegardés et archivés sur la plateforme via une commande de demande de sauvegarde à partir de l'url de l'entrepôt[1].
- Le code source est sauvegardé et l'archive lui attribue un identifiant standardisé SWHID qui permet de pointer sur une version précise du code source[2].

Stratégie 2 : dépôt avec métadonnées et obtention d'un DOI citable (Zenodo)

- Les codes sources gérés sur le Gitlab d'Huma-Num pourront être déposés manuellement sur Zenodo, afin d'obtenir un DOI et une description par des métadonnées.
- Les codes sources gérés sur Github peuvent être déposés automatiquement à partir de chaque version formalisée dans la plateforme (tags).

Stratégie 3 : dépôts via HAL et métadonnées, obtention d'un identifiant citable *SWHID* et modérés

- Les codes sources gérés sur le Gitlab d'Huma-Num pourront également être déposés manuellement sur l'archive ouverte HAL avec ajout de 3 fichiers texte (README, AUTHORS et LICENSE) et une fiche de métadonnées minimale[3]. Le dépôt via HAL est modéré, ce qui apporte une forte visibilité au code déposé et une citation fiabilisée.

23.1.2 A.2 Données du cluster de données et référentiels d'autorité

Les données du portail et des référentiels d'autorité sont partagées par les API et web services qui permettent d'extraire les données en totalité.

Des exports au format RDF de chaque référentiel seront réalisés à intervalles réguliers et déposés sur Zenodo. Ils feront l'objet de data papers dans des revues (par exemple *Humanités numériques*, *Journal of Open Humanities Data*, etc.).

23.1.3 A.3 Données sur Nakala

Même si ce n'est pas spécifiquement prévu au stade de la version initiale du PGD, il est possible que certaines données de **Biblissima+** soient également déposées sur Nakala, soit pour bénéficier de ses fonctionnalités particulières de publication de collections (Nakala Press) soit pour produire des collections cohérentes en lien avec les équipes partenaires qui auraient fait le choix de cette plateforme (notamment pour certains corpus de textes ou de partitions encodés en TEI). Il est à noter que les données déposées dans Nakala peuvent faire l'objet d'un projet de préservation avancée dans le cadre d'une convention de partenariat passé entre Huma-Num et le CINES[4]. Ce service est accessible pour des corpus sélectionnés par le comité de liaison[5].

23.2 B/ Méthodes et outils nécessaires pour accéder aux données et les utiliser

Les outils logiciels nécessaires pour accéder aux données et les utiliser varient selon les formats choisis. Pour le détail, se reporter aux tableaux synthétiques de la partie précédente. En règle générale, les données textuelles formalisées au format **XML** ou **RDF** sont lisibles avec un simple éditeur de texte basique. En revanche le volume de certains fichiers peut nécessiter l'emploi d'un logiciel spécifique optimisant la consultation ou l'interrogation des données (par exemple un triplestore pour les dumps RDF, ou un moteur de base de données de type **BaseX** ou **eXist-db** pour les fichiers **XML** volumineux).

23.3 C/ Attribution d'identifiants pérennes uniques

Les stratégies de dépôt des codes sources du périmètre 1 de **Biblissima+** permettent de combiner les avantages des deux types d'identifiants pérennes et uniques offerts par l'état de l'art en matière d'identification de codes sources et de données : les identifiants extrinsèques et les identifiants intrinsèques. Les premiers utilisent un registre pour conserver la correspondance entre l'identifiant et l'objet identifié. Les seconds reposent sur un accord sur la méthode à employer pour les calculer. Ils peuvent donc se passer de registres et d'une autorité garante.[6]

23.3.1 Identifiants pérennes et uniques des codes sources

Les codes sources ou jeux de données déposés dans Zenodo ou Nakala se voient automatiquement attribuer un identifiant pérenne unique extrinsèque de type **DOI**. Les codes sources archivés dans Software Heritage bénéficient d'un identifiant pérenne intrinsèque *SWHID* doté des fonctionnalités nécessaires à la citation de code source, tout en restant indépendant de l'implémentation technique de la gestion du code source (contrairement aux identifiants apportés par les plateformes **Github** ou **Gitlab** qui restent techniquement dépendants des choix techniques de ces plateformes).

23.3.2 Identifiants pérennes des entités des référentiels gérés via Wikibase

Les identifiants des entités des référentiels d'autorité sont construits comme les entités **Wikidata** sous la forme d'identifiants numériques préfixés par la lettre Q (ex. Q2987). Ils sont générés par l'instance Wikibase administrée par l'équipe portail. Si un doublon est repéré parmi les entités des référentiels, la plateforme permet de fusionner les deux entités et une reconduction des identifiants concernés est assurée de façon automatique. Cette fusion peut se faire aussi bien manuellement via l'interface de la plateforme, qu'automatiquement via l'API de **Wikibase**.

23.3.3 Identifiants des données du portail

Un système d'identifiants pérennes **ARK** a été mis en place pour les pages du portail **Biblissima+.** Ces URL **ARK** sont basées sur des identifiants opaques alphanumériques (reposant sur l'algorithme **SHA1**) générés de façon automatique lors de la phase de traitement des entités.

Ces identifiants ne sont donc pas gérés par un logiciel spécifiquement dédié à la gestion d'**ARK** (ex. Noid) et n'implémentent pas toutes les fonctionnalités liées aux ARK (qualificatifs, inflexion pour accéder aux métadonnées). Ils ne s'appuient pas non plus sur un résolveur **ARK** externe de type N2T.net : ainsi le logiciel **Cubicweb**, qui propulse le portail **Biblissima+**, agit comme le résolveur local des **ARK** Biblissima. Le coût de mise en place d'une infrastructure complète de gestion des **ARK** a été jugé trop important dans le cadre d'un projet d'ÉquipEx d'une durée limitée. Cependant, il est envisagé de verser à la fin du projet tous les identifiants **ARK** Biblissima dans le résolveur **ARK** global N2T.net, voire de les migrer vers une autre solution de type DOI, de sorte que leur maintenance puisse être transférée à une autre entité institutionnelle (par exemple le porteur **Établissement Public Campus Condorcet**). En conséquence l'utilisation des identifiants Biblissima au sein des ressources et corpus fournis par les équipes partenaires constituent un enrichissement pérenne et d'interopérabilité entre ressources valables à long terme et au-delà du contexte de **Biblissima+**.

Si un doublon est repéré parmi les pages du portail, la fusion de l'ancien identifiant **ARK** avec le nouveau est faite manuellement. La redirection de l'ancienne URL vers la nouvelle s'appuie sur une table de redirection maintenue par l'équipe et paramétrée au niveau du serveur Web.

---

1. Utilisation possible d'une API pour automatiser le processus. ↩
2. Pour plus d'informations sur le schéma d'identifiants et la plateforme, voir la documentation : https://docs.softwareheritage.org/ ↩
3. https://www.softwareheritage.org/2019/11/28/saving-and-referencing-research-software-in-software-heritage/?lang=fr ↩
4. cf. https://documentation.huma-num.fr/nakala-faq/#lors-de-la-demande-de-preservation-a-long-term ↩
5. https://documentation.huma-num.fr/partenariat-hn-cines/ ↩
6. D'après https://bbf.enssib.fr/consulter/bbf-2021-00-0000-002 ↩

24 Ressources et responsabilités

24.1 A/ Responsable de la gestion des données

La gestion des données au sein du périmètre P1 est réalisée par l'équipe portail, sous la responsabilité de l'**Établissement public Campus Condorcet**. Elle est suivie par la directrice adjointe et coordinatrice du volet A « Infrastructure numérique ».

La gestion des données au sein des périmètres P2 et P3 est placée sous la responsabilité des équipes partenaires et de leurs établissements de tutelle.

Voir aussi l'accord de consortium **Biblissima+** signé en avril 2023.

Activité	Responsabilités fonctionnelles
Saisie des données	Équipe portail
Production des métadonnées	Équipe portail
Qualité des métadonnées	Équipe portail
Qualité des données	Équipe portail et responsables scientifiques et techniques des livrables au sein du partenariat
Stockage et sauvegarde	Équipe portail
Partage et archivage des données	Équipe portail
Rédaction du PGD et mise à jour du PGD	Directrice adjointe coordinatrice du volet A
Validation du PGD	Responsable scientifique et technique de l'ÉquipEx

24.2 B/ Ressources permettant de s’assurer que les données seront FAIR

Les plateformes utilisées pour le stockage, le partage et l'archivage n’impliquent pas de coûts financiers pour les dépôts ne dépassant pas un volume de données de 50 Go (**Zenodo**).

Sur la base de l’expérience acquise dans le cadre de Biblissima 1, le temps de travail cumulé nécessaire pour assurer la gestion des données d’une ressource intégrée au sein du cluster de données est estimé d’une durée de 1 à 2 personnes.jour.

V. Annexes

25 Historique des révisions et validations

Version	Date	Modifié par	Commentaire
1	24/03/2022	E. Morlock	Première structuration et ébauche à partir du modèle ANR.
1.1	28/03/2022	E. Morlock	Premier brouillon soumis à l'équipe portail pour rectifications et compléments (sur l'ensemble du plan hors annexes et études de cas).
1.2	01/04/2022	Régis Robineau, Eduard Frunzeanu	Compléments, ajouts, corrections de 0.1
1.3	02/04/2022	E. Morlock	Ajout des éléments périmètre 2 (tableaux de synthèse).
1.4	05/04/2022	M.-A. Avenel, F. Bougard, A.-M Turcan-Verkerk	Corrections et ajouts.
1.5	15/04/2022	E. Morlock	Finalisation et mise en page.
1.6	25/04/2022	E. Morlock	Intégration des corrections indiquées par les responsables scientifiques et techniques de livrables. Mise en page finale.
1.7	26/04/2022	E. Morlock	Intégration des corrections de A.-M Turcan-Verkerk et mise à jour du tableau 3 de la partie vue d'ensemble.
			Version à 6 mois transmise à l'ANR.
1.8	15/09/2022	E. Morlock	Corrections ortho-typographiques (pp. 4, 8, 17,20,22,50, 56, 61, 62, …) ; numérotation des annexes ; Gloss-e : ajout équipe LEM p. 32 ; Edition des Gloses : ajout de la licence ; Equipes p. 37 ; ajout hébergement note 20 p. 37, réécriture du

Version	Date	Modifié par	Commentaire
			paragraphe A.3 données sur Nakala p. 61, DOI Nakala p. 62, correction des liens vers les tutoriels Zenodo p. 62.
1.9	02/10/2022	E. Morlock, R. Robineau	Version numérique dans le système de publication Mkdocs pour publication en ligne à l'adresse : https://dmp.biblissima.fr
2	26/10/2023	E. Morlock, R. Robineau, Eduard Frunzeanu, K. Bois	Version 2 envoyée à l'ANR sous forme de fichier pdf, déposé dans Zenodo et mis en ligne dans le site en ligne mkdocs sur dmp.biblissima.fr.

26 Abréviations, sigles et acronymes

`ABES`

Agence Bibliographique de l'Enseignement Supérieur.

`ANR`

Agence nationale de la Recherche – https://anr.fr/fr/.

`API`

Application Programming Interface (interface au sein d'une application logicielle permettant à d'autres applications d'accéder à une sélection de fonctionnalités et de transférer des données dans les deux sens).

`ARGOS`

Plateforme d'aide à la rédaction collaborative de plans de gestion de données (Data Management Plans) du projet OpenAIRE intégrée aux services de l'initiative (EOSC) de la Commission européenne.

`ARK`

Archival Resource Key (format d'identifiant créé par la California Digital Library fournissant un mécanisme d'identification pérenne des objets.

`B1`

Biblissima, 1ère période de financement de l'ÉquipEx, d'octobre 2012 à décembre 2019, prolongée jusqu'en décembre 2021 (référence : ANR-11EQPX0007).

`B+`

Biblissima+, période de financement actuelle, de novembre 2021 à juin 2029 (référence : ANR-21-ESRE-0005).

`CC`

Creative Commons (contrats de licences ouvertes permettant aux auteurs d'autoriser des modalités d'exploitation de leurs œuvres à partir d'options prédéfinies portant sur l'attribution, l'utilisation commerciale, le partage et la modification).

`CIDOC-CRM`

Conceptual Reference Model (CRM) developed by the International Committee for Documentation (CIDOC) of the International Council of Museums (ICOM).

`CINES` v

Centre Informatique National de l'Enseignement Supérieur – https://www.cines.fr/.

`Codemeta`

Standard d'échanges de métadonnées de logiciels entre entrepôts – https://codemeta.github.io/.

`CPER`

Contrat de Plan État Région.

`DMP-OPIDOR`

Plateforme collaborative d'aide à la rédaction de plans de gestion de données et de logiciels mise à disposition par l'INIST-CNRS, accessible à la communauté scientifique de l'ESR et à ses partenaires français ou étrangers – https://dmp.opidor.fr/.

`DOI`

Digital Object Identifier – https://www.doi.org

`DORANUM`

DOnnées de la Recherche Apprentissage NUMérique – ressources d'auto-formation sur la gestion et le partage des données de la recherche (Réseau des Unités régionales de formation à l'information scientifique et technique, l'INIST-CNRS et représentants de la communauté de l'enseignement supérieur et de la recherche). https://doranum.fr/.

`EAD`

Encoded Archival Description (Description archivistique encodée) – https://www.loc.gov/ead/.

`FAIR`

Findable, Accessible, Interoperable, Reusable (ou en français : Facilement trouvable, Accessible, Interopérable, Réutilisable) – https://www.go-fair.org/fair-principles/.

`FACILE`

Service de validation de formats – https://facile.cines.fr/

`FNE`

Fichier national d'entités (piloté par l'ABES).

`HAL`

Archive ouverte pluridisciplinaire destinée au dépôt et à la diffusion de publications scientifiques – https://hal.archives-ouvertes.fr/.

`HTR`

Handwritten Text Recognition (transcription automatisée de sources manuscrites).

`Huma-Num`

Infrastructure de Recherche Huma-Num (IR*, appelée TGIR – très grande infrastructure de recherche – dans les précédentes éditions de la Feuille de route nationale) – https://www.huma-num.fr.

`IIIF`

International Image Interoperability Framework™ – ensemble de standards qui définissent un cadre d'interopérabilité pour la diffusion des images numériques sur le web – https://iiif.io/ et https://iiif.biblissima.fr.

`INRIA`

Institut national de recherche en sciences et technologies du numérique – https://www.inria.fr/.

`ISMI`

International Standard Manuscript Identifier (registre électronique des identifiants des livres manuscrits).

`ISNI`

International Standard Name Identifier (Code international normalisé des noms).

`NOID`

Nice Opaque Identifier (codes alphanumériques fournissant un mécanisme d'identification des objets, numériques ou non numériques, ne portant pas de signification).

`OCR`

Optical Character Recognition (Reconnaissance optique de caractères).

`OpenAIRE`

Open Access Infrastructure for Research in Europe – Projet financé par la Commission européenne visant à diffuser en accès ouvert les publications et les données scientifiques issus des travaux des différents projets européens.

`OpenEdition`

Portail de ressources électroniques en sciences humaines et sociales, comprenant 4 plateformes dédiées respectivement aux livres, aux revues, aux blogs de recherche et aux annonces scientifiques.

`OPERAS`

Infrastructure de recherche ayant pour mission de soutenir la communication scientifique ouverte en sciences humaines et sociales au sein de l'Espace européen de la recherche (EER).

`OpenRefine`

outil libre d'extraction, de nettoyage et d'alignement de données – https://openrefine.org.

`PGD`

Plan de gestion des données (Data Management Plan).

`PID`

Persistent Identifier (Identifiant pérenne).

`RDA`

Research Data Alliance (organisation internationale développant des activités communautaires pour favoriser le partage ouvert des données et la réutilisation des données https://www.ouvrirlascience.fr/research-data-alliance-rda/).

`RGPD`

Règlement Général sur la Protection des Données.

`SF`

Software Heritage, plateforme d'archivage pérenne de logiciels développée dans le cadre d'une organisation à but non lucratif soutenue par plusieurs partenaires institutionnels, lancée en 2016 par l'INRIA et soutenue par l'UNESCO – https://www.softwareheritage.org.

`SUDOC`

Système universitaire de documentation (piloté par l'ABES).

`SWHID`

Software Heritage identifier (identifiant pérenne utilisé par la plateforme d'archivage des codes sources Software Heritage) – https://www.softwareheritage.org/

`TEI`

Text Encoding Initiative – Standard d'encodage XML utilisé notamment pour la création et l'exploitation d'éditions électroniques adaptées aux besoins des chercheurs en sciences humaines et sociales, comme les éditions de sources historiques, de manuscrits, de documents d'archives, inscriptions antiques, etc.

`TRIPLE`

Transforming Research through Innovative Practices for Linked Interdisciplinary Exploration - Service de l'infrastructure de recherche OPERAS visant à offrir une plateforme multilingue et multiculturelle pour la découverte de projets, publications et données en sciences humaines et sociales – https://project.gotriple.eu/.

V.I Méthodologie

27 Création du PGD V1 (avril 2022)

La première version du PGD a été préparée par une « enquête » auprès des équipes partenaires en février et mars 2022. Un modèle de PGD simplifié, mis en forme sous forme de tableur, a été envoyé aux membres du comité de direction afin qu'ils le transmettent aux chercheurs, enseignants-chercheurs ou ingénieurs responsables de livrables de leurs équipes ou unités. La demande était de remplir un tableau par livrable et de dupliquer la grille d'analyse afin de renseigner un onglet par jeu de données identifié. Par exemple, une édition de corpus en TEI peut donner lieu à une collection de fichiers XML, une collection d'images fac similaires, une interface de saisie et un site de publication web… Les modalités de gestion, de partage et d'archivage peuvent faire appel à des procédures, des outils ou des licences de diffusion très hétérogènes. Les responsables de livrables ont renvoyé les tableaux renseignés entre le 26 février et le 20 mars 2022. Le taux de réponse par rapport à l'ensemble des livrables concernés par la question des données ou des codes sources est d'un tiers environ. Ce taux peut s'expliquer par le fait que les porteurs d'opérations ne débutant pas avant 3 ou 4 ans ne se soient pas sentis concernés.

Ce recueil d'informations avait aussi pour finalité de sensibiliser au PGD, en montrant notamment que les questions sont plus d'ordre organisationnel et stratégique que purement technique. Les informations renseignées dans les tableaux ont été utilisées pour les tableaux synthétiques de la partie « vue d'ensemble ». Elles serviront également plus tard à l'équipe portail pour la planification des collaborations avec les équipes scientifiques.

Le PGD principal a été rédigé par la directrice adjointe coordinatrice du volet A « Infrastructure numérique » et l'équipe portail. Les membres du bureau exécutif ont ensuite revu et validé une première version de la rédaction. Le document a ensuite été soumis pour commentaire et avis à l'ensemble des membres du projet du 15 au 20 avril 2022. La version envoyée à l'ANR prend en compte les remarques formulées dans ce cadre.

27.1 Questionnaire de recueil d'informations (périmètre P2)

Description du jeu de données	vos réponses dans cette colonne (voir l'exemple)
Période de début des travaux prévue (en indiquant une tranche annuelle ou semestrielle par rapport à la durée totale du financement ANR, par ex. T18, T60…)	
Période de livraison prévue (idem)	
Modalités d'utilisation des référentiels Biblissima	

Caractérisation des données collectées (produites ou réutilisées)	vos réponses dans cette colonne (voir l'exemple)
Type de provenance (Création de nouvelles données / réutilisation et transformation de données existantes).	
Type de données (textuelles / numériques / images / vidéo / médias divers / de simulation / code informatiques).	
Format(s) des données et Standard(s) utilisés (exprimés à l'aide de l'extension du nom de fichier .Txt, .pdf, .csv) - préciser s'il y a un encodage standard (XML/TEI, EAD)…	
Informations sur la volumétrie (exprimés en espace de stockage requis (octets), et/ou en quantités d'objets, de fichiers, de lignes, et colonnes).	

Métadonnées et documentation	vos réponses dans cette colonne (voir l'exemple)
Comment les métadonnées des fichiers regroupés dans le jeu de données sont-elles produites ?	
Standard(s) ou schéma(s) de métadonnées utilisées pour renseigner les métadonnées (par exemple Dublin Core, TEI, EAD, Datacite…).	
Y a-t-il des éléments de documentation indispensables pour permettre la réutilisation des données (méthodologie de collecte, procédures et méthodes d'analyse, définition des variables et des unités de mesure…) ?	

Stockage et sauvegarde des données et métadonnées pendant le projet	vos réponses dans cette colonne (voir l'exemple)
Type d'hébergement et lieu de stockage au cours du processus de recherche et d'élaboration du livrable (préciser la fréquence de sauvegarde ou le plan de sauvegarde s'il existe).	
Qui aura accès aux données du livrable au cours du processus de recherche (et comment l'accès aux données est contrôlé, en particulier dans le cadre de recherches menées en collaboration).	
En cas de données sensibles (par exemple données à caractère personnel, politiquement sensibles des informations ou secrets commerciaux), décrire les principaux risques et la façon dont ils seront gérés pour le livrable.	

Titularité, exigences légales et éthiques	vos réponses dans cette colonne (voir l'exemple)
Qui aura le droit de contrôler l'accès aux données du livrable ? (lister tous les partenaires le cas échéant). Indiquer si les droits de propriété intellectuelle sont affectés.	
Du matériel protégé par des droits spécifiques sera-t-il utilisé au cours du projet (ex : données personnelles, bases de données…) ?	
Indiquer s'il y a des restrictions sur la réutilisation de données fournies par des tiers et en expliquer les raisons le cas échéant (par exemple données soumises à des droits de propriété intellectuelle, de confidentialité contractuelle, de sécurité…)	

Partage, conditions de réutilisation et DOI	vos réponses dans cette colonne (voir l'exemple)
Nom du ou des entrepôt(s) dans lequel(s) une copie du jeu de donnée sera déposée. Par exemple : Nakala, Zenodo, etc. Pour les codes logiciels : Github, Gitlab (préciser l'institution hébergeante). Si l'entrepôt n'attribue pas d'identifiant pérenne, préciser comment celui-ci est obtenu. A défaut, par quels autres moyens le jeu de données pourra-t-il être retrouvé et partagé ?	
En cas d'interdit au partage ou d'embargo, indiquez les raisons et les durées (publication, protection de la propriété intellectuelle, dépôt de brevet…)	
Quelles méthodes ou quels outils logiciels seront nécessaires pour accéder aux données et les utiliser ?	
Quelle licence de réutilisation sera appliquée au jeu de données ? (Creative Commons, Licence ouverte, Open database licence, etc. cf. https://www.data.gouv.fr/fr/pages/legal/licences/).	
Un identifiant unique et pérenne sera-t-il attribué aux données publiées en ligne ? Si oui, lequel ? Si non, quel autre type d'identifiant sera attribué (URL, identifiant local, pas d'identifiant…) ?	

Conservation à long terme	vos réponses dans cette colonne (voir l'exemple)
Le jeu de données est-il concerné par la conservation à long terme ? Si oui, indiquer les principes et les procédures selon lesquelles les données seront sélectionnées.	
Quelle plateforme est envisagée pour la conservation à long terme ? Précisez le nom de l'institution prenant en charge les coûts. S'il s'agit d'un archivage au CINES, précisez qui sera en charge de définir le workflow des échanges de données.	
Indiquez la volumétrie estimée pour l'archivage à long terme.	

Rôles, Responsabilités & coûts	vos réponses dans cette colonne (voir l'exemple)
Responsable de la gestion des données (stockage, partage, archivage…)	
Responsable de la rédaction et mise à jour du PGD du livrable (qui sera à rédiger entre T6 et T12).	
Quelles seront les ressources (budget et temps alloués) dédiées à la gestion des données permettant de s'assurer que les données seront FAIR ? (FAIR : Facile à trouver, Accessible, Interopérable, Réutilisable cf. https://doranum.fr/enjeux-benefices/principes-fair/).	

27.2 Synthèse des réponses

Le découpage en lignes de financement associées à des « livrables » réalisé pour définir le calendrier des versements financiers[1] comporte environ 125 livrables produisant des jeux de données, des codes sources ou des méthodes nécessitant l'établissement d'un plan de gestion des données. Les réponses reçues couvrent 47 livrables, ce qui correspond à un taux de réponse de 37 %. Le nombre de réponses ne permet pas une analyse très poussée. Quelques observations peuvent néanmoins être tirées de ces documents.

27.2.1 Description du jeu de données

Les descriptions complètent les informations données en 2020 lors du montage de la proposition[2]. La question utile pour l'équipe technique de **Biblissima+** porte sur l'utilisation des référentiels au sein du projet. Elle facilitera l'identification des producteurs de données avec qui l'équipe portail devra travailler plus directement.

27.2.2 Caractérisation des données collectées

- Les livrables portent autant sur la fourniture de nouvelles données que sur la reprise de données existantes. La plupart du temps, il s'agit de partir d'une base existante (fichier TEI, code logiciel) et de l'enrichir ou la développer. Il n'y a pas de cas de réutilisation « telle quelle », sans modification, dans les exemples reçus. La catégorie de données « transformées » est à ajouter à la typologie pour les futures versions du PGD.
- Les formats utilisés sont très variés mais correspondent à l'état de l'art ainsi qu'aux bonnes pratiques des communautés impliquées. Pour les codes sources sont principalement représentés les langages de manipulation du format XML, les langages de programmation Python, R, et les langages du web comme Javascript, HTML/CSS et Json, Pour les images, les formats TIFF, SVG et Jpeg sont mentionnés ainsi que le protocole IIIF. Les standards liés aux textes structurés les plus représentés sont TEI, EAD et RDF.
- Peu d'informations sont données sur la volumétrie, probablement du fait des faibles volumes occupés par les données textuelles et les codes logiciels. Un livrable parle de 1 To de données pour une collection d'images au format TIFF.

27.2.3 Métadonnées et documentation

- Les fiches de métadonnées des dépôts seront la plupart du temps remplies manuellement. Les standards utilisés dépendent de la plateforme de dépôt choisie : Datacite s'il s'agit de Zenodo, Dublin Core en ce qui concerne Nakala.

27.2.4 Stockage et sauvegarde

- La plupart des répondants bénéficient d'un environnement de travail doté d'espaces serveur ou Cloud apportés par l'un de leurs établissements tutelles. Quand ce n'est pas le cas, ils recourent aux services d'Huma-Num (Sharedocs, hébergement web…).
- L'accès aux données est réservé aux membres de l'équipe pendant les travaux puis rendu public à la fin des travaux.

27.2.5 Titularité des droits d'auteur, exigences légales et éthiques

- La question de la propriété intellectuelle des contenus produits par les chercheurs ne se pose que dans certains cas particuliers où les travaux sont centrés sur l'annotation critique de corpus (mais certains projets du même type ouvrent leurs données avec les licences CC BY qui garantissent la mention d'attribution aux auteurs).
- Les répondants n'ont pas déclaré de traitements de données sensibles ou personnelles.

27.2.6 Partage conditions de réutilisation et DOI

- La question du choix de la plateforme de dépôt ne semble pas vraiment difficile pour une grande partie des répondants, tandis que d'autres indiquent « je ne sais pas » en réponse sur ce point. Les réponses ne citent que les entrepôts génériques Zenodo et Nakala. Le portail Persée gère son propre entrepôt.
- Le partage des données ouvert ne fait pas débat dans la grande majorité des cas. La licence choisie est soit Creative Commons, soit la licence très permissive recommandée par la loi *Pour une république numérique* de 2016 (Licence ouverte Etalab 2.0). Les raisons de choisir l'une plutôt que l'autre ne ressortent pas nettement des réponses : le même type de ressource optera suivant les cas pour l'une plutôt que pour l'autre, plus par habitude qu'en raison d'une analyse de leurs différences, semble-t-il. Souvent les répondants citent les deux types de licences en précisant que la décision sera prise plus tard. L'usage des options restrictives "partage à l'identique" (*Share Alike - SA*), "Pas de modification" (*No Derivatives* ou *ND*) ou "pas d'utilisation commerciale" (*Non commercial - NC*) n'est pas justifiée et pourrait être débattue dans certains cas. Les clusters ont certainement un rôle à jouer pour faciliter le choix des licences en formulant des recommandations adaptées au contexte spécifique.

27.2.7 Conservation à long terme

- La question de la conservation ou préservation à long terme ainsi que de l'archivage pérenne ne semble pas entièrement comprise – ce qui n'est pas une surprise étant donné la complexité du domaine de la pérennisation numérique. Pour certains répondants, l'utilisation des espaces d'hébergement web ou de la plateforme Nakala vaut archivage pérenne. Le fait que l'archivage pérenne au CINES avec Nakala n'est qu'une possibilité soumise à audit et à l'établissement d'une convention avec le CINES ne semble pas connue. Seule la plateforme Persée cite la norme OAIS et la plateforme précise utilisée par le CINES pour l'archivage pérenne PAC.
- Pour les codes sources, aucun répondant ne cite explicitement l'archive Software Heritage. La possibilité d'utiliser conjointement Github et Zenodo pour archiver des versions majeures citables par DOI est mentionnée par un répondant. Pour le domaine TEI, le format ODD n'est pas cité comme un format intéressant pour l'archivage de la TEI, même s'il figure dans la liste des standards utilisés par ailleurs

27.2.8 Rôles responsabilités et coûts

- Les responsabilités dans la gestion des données telles que les répondants les présentent sont confiées à trois types d'acteurs : les responsables scientifiques et techniques, les informaticiens ou les post-doctorants et ingénieurs recrutés grâce à l'aide financière apportée par **Biblissima+**. Il n'y a pas de coûts financiers identifiés, probablement grâce aux conditions de gratuité offertes par les infrastructures mutualisées accessibles à la communauté académique en France. En ce qui concerne la charge de travail, celle-ci est estimée la plupart du temps à une durée de 2 à 5 personnes jours à l'échelle d'un livrable ou d'une personne jour par an et par jeu de données.

---

1. cf. l'annexe Livrables de **Biblissima+** donnant lieu à un versement financier ↩
2. Document rédigé par les équipes partenaires qui présente en détail l'infrastructure numérique envisagée (livre blanc téléchargeable depuis la page : https://projet.biblissima.fr/fr/projet/presentation) et dans la communauté Zenodo Biblissima+ via l'identifiant DOI 10.5281/zenodo.6611721 ↩

28 Mise à jour du PGD V2 (octobre 2023)

Un questionnaire a été envoyé sur la liste de diffusion “biblissima-info-membres” le 3/10/2023 via l’outil en ligne framaforms[1].

Ce questionnaire était adressé plus spécifiquement aux responsables scientifiques et techniques de livrables ainsi qu’aux porteur.e.s d’un projet laureat de l’appel à projet annuel. IL leur était demandé de vérifier les tableaux de synthèse de la version 1 du PGD et d’indiquer les modifications souhaitées.

Les questions suivantes portaient sur le plan particulier établi pour le ou les livrables financées par Biblissima : existence d’un PGD, souhait de le communiquer ou non à l’équipe technique de Biblissima+, la licence de diffusion associés.

28.1 Réponses au questionnaire

12 réponses ont été reçues au 23/10/2023 (11 via le questionnaire + 1 par mail). Une L’invitation à remplir le questionnaire a été envoyée à la liste de diffusion `<biblissima-info-membres@listes.campus-condorcet.fr>` (155 abonnés).

question	oui	non
Souhaitez-vous demander des corrections ou ajouts ?	5	6
Avez-vous rédigé un PGD pour votre projet financé ou co-financé par Biblissima+	4	7
Si vous avez répondu “Oui” à la question précédente, souhaitez-vous le porter ce document à la connaissance du bureau exécutif et de l’équipe technique de Biblissima+ ?	1	3
Si vous avez répondu “Oui” à la question précédente, le lien ou le fichier peuvent-il être diffusés ?	0	1

28.2 Enquête par mail

Questionnaire - envoyé par mail le 3/10/2023 (liste biblissima-info-membres@listes.campus-condorcet.fr)

Bonjour à toutes et tous,

Ce mail s'adresse plus particulièrement à tous les responsables scientifiques et techniques de livrables financés par Biblissima+. Les lauréats de l'appel à projets sont également concernés.

La mise à jour du plan de gestion des données (PGD) de Biblissima+, soit la version 2, est à envoyer à l'ANR avant le 26 octobre prochain.

Le document est en cours de mise en forme sur un site en ligne, sur le modèle du vademecum du programme (cf. https://doc.biblissima.fr/vademecum-biblissima/). Cette mise en ligne facilitera la consultation, les mises à jour ultérieures et permettra des liens de citation directs. Elle rendra en outre les corrections, ajouts ou suppressions traçables.

Je vous écris aujourd'hui pour vous demander de lire les parties vous concernant, vous, votre équipe de rattachement ou votre projet, afin de m'indiquer quelles modifications à apporter, si possible avant le 16 octobre 2023.

J'ai bien conscience que le délai est très cours et vous prie de m'en excuser.

Cependant, les informations demandées sont très limitées : elles portent seulement sur les tableaux de synthèse donnant une vue d'ensemble des données produites. Il faut juste vérifier l'exactitude des informations, compléter les manques – lors qu'est indiqué "non connu" par exemple, ou corriger les éventuelles erreurs.

Ce questionnaire rapide (3 à 7 questions selon les cas) vous permettra de répondre en quelques minutes après avoir consulté le document.

En réalité, il vous suffit de télécharger le pdf, de faire une recherche sur votre nom, celui de votre équipe ou celui de votre projet. Les tableaux de synthèse n'affichent que quelques colonnes. Cette vérification qui vous est demandé peut donc être très rapide.

Je vous en serais très reconnaissante.

Voici le lien : https://framaforms.org/questionnaire-rapide-pgd-v2-1696272639

Quelques informations sur la mise à jour des autres parties du PGD : - La mise à jour des principes directeurs n'évoluera qu'à la marge. - Des modifications de détails seront également apportées à la partie PGD détaillée, centrée uniquement sur l'infrastructure numérique gérée par l'équipe technique Biblissima.

Bien entendu, je reste à votre disposition pour toute question concernant la démarche de gestion des données et cette mise à jour.

Bien cordialement,

Emmanuelle Morlock Dir adj. de Biblissima+ (suivi du volet A)

---

1. Export au format txt consultable via ce fichier : webform-form1-799455.txt) ↩

V.II Livrables

29 Ressources financées dans le premier ÉquipEx Biblissima

À l'exception des projets terminés au cours de Biblissima (comme Comparatio, Esprit des livres, Manuscripta medica, RegeCart), toutes ces bases de données sont vivantes et connaissent un développement ininterrompu, grâce aux financements Biblissima (projets d'origine, nouveaux projets, projets partenariaux : la concentration des données produites dans les bases existantes a constamment été favorisée), grâce aux financements récurrents des établissements porteurs et grâce à la mise en place de nouveaux projets et partenariats. Nombre de ces ressources poursuivront leur collaboration avec **Biblissima+**, en particulier pour la mise en commun des référentiels et pour la mise en place d'une automatisation des mises à jour.

La version courante du portail Biblissima intègre pour le moment des jeux de données issus de 19 sources, ce qui représentait près de 650 000 pages fin 2021.

Voir aussi la page : https://portail.biblissima.fr/a-propos

#	Ressource	Description
1	Bibale (IRHT)	Données de provenance des bibliothèques françaises : histoire de la transmission des livres et manuscrits et imprimés par l'étude des collections anciennes et modernes et de leurs possesseurs (version intégrée : version 1, version actuelle : version 2). – Devenue le hub interopérable de l'IRHT pour toutes les informations sur les personnes.
2	Bibliotheques françoises de La Croix du Maine et de Du Verdier (1584 et 1585) (CESR)	Edition et base de données en TEI (données bibliographiques et prosopographiques)
3	Books within Books (EPHE et partenaires internationaux)	Base sur les fragments de manuscrits hébreux.
4	BUDE (IRHT puis CESR)	Base sur les mains d'humanistes et les correspondances d'érudits de la Renaissance.
5	Collecta (Ecole du Louvre puis IRHT)	Base de données en ligne issue de la documentation extraordinaire accumulée par l'érudit François Roger de Gaignières (1642-1715)
6	Comparatio (IRHT)	Base de données sur le chant liturgique
7	Projets CR2I et CRIICO (CESR)	Rétroconversion des catalogues imprimés d'incunables des bibliothèques de France
8	Esprit des livres (Ecole nationale des chartes)	Catalogues de vente de bibliothèques de l'époque moderne et en particulier les manuscrits anciens passés en vente.
9	Europeana Regia	Base achevée en 2012 et dont le site est obsolète. L'intégration des données au portail Biblissima sauve les données et leur structuration.
10	Initiale (IRHT)	Base de notices iconographiques, notices de possesseurs, bibliographie.

#	Ressource	Description
11	Jonas (IRHT)	Base mettant à disposition toute la documentation de l'IRHT sur les textes romans.
12	Mandragore (BnF)	Base d'enluminures décrivant les illustrations des manuscrits du département des Manuscrits et de la Bibliothèque de l'Arsenal. Mise en interopérabilité sur le portail Biblissima et une consultation plus aisée des numérisations de la BnF grâce au visualiseur Mirador.
13	Manuscripta medica (EPHE et CIHAM)	Base décrivant l'ensemble des manuscrits médicaux des bibliothèques publiques de France.
14	Medium (IRHT)	Base de données des manuscrits numérisés par l'IRHT, qui sert de référentiel de cotes pour l'ensemble du laboratoire.
15	Pinakes (IRHT)	Base regroupant toute l'information disponible sur les manuscrits grecs conservés et est la base de référence du domaine
16	Reliures (BnF)	Base et son schéma TEI mis à disposition de tous via le site web du projet.
17	RegeCart (IRHT)	Base de données qui donne accès à l'analyse de 571 cartulaires, cartulaires-chroniques ou bullaires. Ces analyses accumulées par la section de diplomatique de l'IRHT sont un formidable gisement de faits liés à des personnes, des lieux, des dates.
18	SourcEncyMe (Nancy puis IRHT)	Corpus en ligne consacré aux encyclopédies médiévales et à leurs sources
19	Anciennes collections de manuscrits grecs (EPHE)	La base n'a pas été développée par Biblissima, qui aurait préféré une mise

#	Ressource	Description
		en ligne des données dans Bibale ou Pinakes, mais son alimentation l'a été. Cette base est la seule qui ne soit pas en ligne et pour laquelle nous n'ayons pas de reporting à jour ; pas de trace ici non plus : https://www.saprat.fr/bases-de-donnees-23.htm).
20	Notices du catalogue de manuscrits classiques latins de la Bibliothèque Vaticane	Test mené avec Persée pour ouvrir le portail Biblissima à la bibliographie en texte intégral (Documents, études et répertoires de l'Institut de Recherche et d'Histoire des Textes, XXI, 5 volumes, disponibles sur Persée) ont été mises en interopérabilité avec les autres ressources du portail, en particulier les bases Pinakes, Bibale, les notices des manuscrits de Heidelberg et les numérisations de la Bibliothèque Vaticane.

30 Livrables de Biblissima+ donnant lieu à versement financier

Le tableau ci-dessous est extrait du document plus complet créé pour organiser le calendrier des versements financiers annuels de l'aide de l'État gérée par l'ANR au titre du programme d'Investissements d'avenir. La numérotation est postérieure et n'a pas de caractère contractuel ou officiel. Elle a principalement pour fonction de faciliter l'identification des jeux de données avec les activités scientifiques et techniques donnant lieu à un versement financier de l'ANR.

Equipe	Intitulé	Référence	Remarque
Bbma	AAP en années 1 à 5	AD_01_CC	-
AOROC	Maintenance du cluster de calcul et de stockage pendant les années de construction (5 ans)	AD_02_CC	-
CESR	Editions de textes TEI-MEI – licences OCR et analyse linguistique	AD_03_UTOURS	-
IRHT	Configuration de pluCo pour Oxygen - 10 licences Oxygen	AD_04_CNRS	-
Bbma	Equipe portail : frais informatiques divers (noms de domaine, licences…)	AD_05_CC	-
Bbma	Communication (flyers, brochures, goodies…) pendant 5 ans	AD_06_CC	-
GED	Stations de numérisation formats A1 et A2	EQ_02_CC	-
Bbma	Accompagnement du déploiement de IIIF dans le réseau des archives dans le cadre de IIIF360 - Equipement informatique	EQ_03_CC	-
Bbma	Equipe portail : matériel informatique pour 3 ingénieurs équipe portail (renouvellement)	EQ_04_CC	-
AOROC	Scanner 3D à mutualiser	EQ_05_EPHE	-
AOROC	Imprimante 3D	EQ_06_EPHE	-
AOROC	Spectromètre portable XRF à mutualiser pour analyse	EQ_07_EPHE	-

Equipe	Intitulé	Référence	Remarque
	physico-chimique		
CESR	Photographie des reliures, poinçons, sceaux, médailles, jetons : appareils photographiques avec objectif macrophotographique	EQ_08_UTOURS	-
CESR	Cartographie et encodage du patrimoine musical : 10 postes informatiques	EQ_09_UTOURS	-
CJM	Un serveur d'inférence et deux serveurs d'entraînement (ressources computationnelles pour les langues à variation graphique)	EQ_10_ENC	-
CJM	Equipement informatique pour l'IR recruté sur ce livrable	EQ_11_ENC	-
CRC	Equipement informatique de l'ingénieur recruté	EQ_12_CNRS	-
CRC	Solution de stockage des données	EQ_13_CNRS	-
CRH	CRH et Editions de l'EHESS : 2 équipements informatiques	EQ_14_EHESS	-
IRHT	Serveur local pour section latine (typologies textuelles)	EQ_15_CNRS	-
IRHT	Equipement informatique 10 postes	EQ_16_CNRS	-
MRSH	Equipement informatique pour 2 ingénieurs	EQ_17_UCAEN	-
Bbma	IIIF 360 : etude et mise en place d'un sparql endpoint et	EQ1_01_CC	-

Equipe	Intitulé	Référence	Remarque
	d'une visionneuse IIIF sur FranceArchives - Equipement informatique		
Bbma	Equipe portail et IIIF360 : missions France et étranger (en part. USA)	MI_01_CC	-
Bbma	Gouvernance : conseil scientifique international, comité de suivi, comité de direction	MI_02_CC	-
Bbma	Journées Biblissima : 2 jours par an (tous les pôles et clusters)	MI_03_CC	-
Bbma	Semaines annuelles des 7 clusters	MI_04_CC	-
CESCM	Missions de terrain, missions formation et compléments pour écoles d'été, missions photographiques	MI_05_CNRS	-
CESR	Missions d'exploration et préparation aap	MI_06_UTOURS	-
CESR	Missions de numérisation ponctuelle en archives	MI_07_UTOURS	-
CESR	Missions pour partenariat avec l'IRHT	MI_08_UTOURS	-
CESR	Missions pour formations TEI et accueil des stagiaires	MI_09_UTOURS	-
CIHAM	Missions pour référentiels valeurs et mesures	MI_10_CNRS	-
CIHAM	Missions et frais pour écoles d'été et formations TEI	MI_11_CNRS	-

Equipe	Intitulé	Référence	Remarque
CIHAM	Environnement d'édition des gloses en collab. avec l'IRHT : missions (préparation du recrutement, puis suivi)	MI_12_ENSLYON	-
CJM	Missions sur les divers livrables	MI_13_ENC	-
CRC	Missions pour échanges avec les partenaires	MI_14_CNRS	-
HiSoMA	Invitations de formateurs internationaux	MI_15_CNRS	-
IRHT	Missions environnements d'édition TEI (interactions avec Caen, Tours et Lyon)	MI_16_CNRS	-
IRHT	Missions réseau international de lexicographes	MI_17_CNRS	-
MRSH / CRAHAM	Frais de mission pendant 5 ans (laboratoire TEI)	MI_18_UCAEN	-
Bbma	AAP en année 6 (= prestations)	PE_01_CC	-
Bbma	Développements informatiques Portail Biblissima, plateforme data.biblissima, site outils	PE_02_CC	-
Bbma	Développements informatiques Portail Biblissima, plateforme data.biblissima, Site outils : évolutions	PE_03_CC	-
IRHT	ISMI : Développement, alimentation, pérennisation	PE_04_CNRS	anciennement PE_04_CC
CJM	Référentiel de noms de lieux : acquisition de données	PE_05_CC	anciennement PE_05_CC

Equipe	Intitulé	Référence	Remarque
CIHAM	Développement de référentiels valeurs et mesures - Numérisations	PE_06_CC	anciennement PE_06_CC
CESR	Fairisation des données extraites des archives 37 : prestation d'indexation, traitement des métadonnées, alignement avec les référentiesl	PE_07_UTOURS	-
IRHT	Numérisation et OCRisation de textes latins sur imprimés anciens	PE_08_CNRS	-
CESCM	La numérisation 3D des inscriptions médiévales : Acquisition de corpus de sources interopérables (photogrammétrie, 3D)	PE_09_CNRS	-
CESCM	Acquisition de corpus de sources interopérables - images de la RMN	PE_10_CNRS	-
CESR	Fairisation des données BVH vers portails et outils Bbma et partenaires	PE_11_UTOURS	-
IRHT	Catalogage automatique des manuscrits numérisés : identification des textes issus de HTR par comparaison avec référentiels textuels (Corpus corporum etc.)	PE_12_CNRS	-
IRHT	Reconnaissance d'entités nommées (cote, noms de personnes, titres d'œuvres) et alignement sur des référentiels	PE_13_CNRS	-
CESR	Répertoire des décors typographiques :	PE_14_UTOURS	-

Equipe	Intitulé	Référence	Remarque
	reconnaissance automatique des formes et des fontes pour l'identification des imprimeurs - dévpt bdd		
SAPRAT	Module héraldique – Développement informatique de l'outil et de l'interface	PE_15_EPHE	-
SAPRAT	Module Sigiscript - Adaptation outil PIM - Développement outil de reconnaissance automatique embarqué	PE_16_EPHE	-
HiSoMA	ATTENTION ce livrable à 12 000 € n'est pas dans l'annexe financière déf de l'ANR (Production audiovisuelle de 4 cours d'auto-formation à l'encodage et à la publication de sources TEI et EpiDoc)	PE_17_CNRS	-
CJM	Chaînage des outils d'édition et d'étude des documents d'archives (cartulaires et chartriers) (B-I.6.2 et B-I.8) : acquisition des données textuelles	PE_18_ENC	-
CESR	Editions de textes TEI-MEI – acquisition de transcriptions (Epistémon)	PE_19_UTOURS	-
CESR	Editions de textes TEI-MEI – encodage	PE_20_UTOURS	-
CESR	Encodage MEI de partitions musicales (40 recueils musicaux)	PE_21_UTOURS	-
CIHAM	Amélioration incrémentielle d'un plugin TEI pour un éditeur XML libre (JEdit), et potentiellement pour d'autres	PE_22_CNRS	-

Equipe	Intitulé	Référence	Remarque
	éditeurs libres, selon les évolutions technologiques, et adaptation du plugin pluCo produit par le PDN de Caen		
CESR	Formations TEI niveaux débutant et avancé	PE_23_UTOURS	-
HiSoMA	Développement d'un connecteur DTS au sein de l'outil TEI Publisher pour l'accès aux fragments des corpus EpiDoc et TEI	PE_24_CNRS	-
Bbma	13% d'ETP pour la direction adjointe volet A pendant la phase de construction	VA_01_CC	-
Bbma	Equipe portail Biblissima+ : responsable (Régis Robineau) : 96 mois	VA_02_EPHE	-
Bbma	Equipe portail Biblissima+: spécialiste référentiels (Eduard Frunzeanu) : 96 mois	VA_03_EPHE	-
Bbma	Equipe portail Biblissima+ : développeur (Kévin Bois) : 96 mois	VA_04_EPHE	-
Bbma	IIIF 360 : etude et mise en place d'un sparql endpoint et d'une visionneuse IIIF sur FranceArchives : 14 mois	VA_05_CC	-
Bbma	Accompagnement numérisation IIIF, en particulier dans les AD : 24 mois	VA_06_CC	-
Persée	Intégration des ID Bbma, travail commun avec l'équipe portail : 24 mois	VA_07_ENSLYON	-

Equipe	Intitulé	Référence	Remarque
IRHT	Littérature critique sur les manuscrits (eScriptorium via Medium) : 11 mois	VA_08_CNRS	-
CRH / Editions de l'EHESS	Développement de l'interopérabilité avec la plateforme Savoirs : 12 mois	VA_09_EHESS	-
GED	Alignement des métadonnées et référentiels : 4 mois	VA_10_CC	-
GED	Adaptation des interfaces	VA_11_CC	-
IRHT	ISMI - Alimentation, développement, pérennisation : 12 mois	VA_12_CNRS	-
IRHT	Référentiels de noms de lieux et de personnes dans les cartulaires médiévaux : 10 mois	VA_13_CNRS	-
CJM	Référentiel de noms de lieux : acquisition de données : 13 mois de vacations	VA_14_ENC	-
CJM	Référentiel de noms de lieux : développement d'une API : 6 mois	VA_15_ENC	-
IRHT	Référentiels pour les manuscrits de l'Orient chrétien et byzantin (grec, syriaque, arabe) : 36 mois + 5 mois	VA_16_CNRS	-
CRAHAM	Enrichissement des référentiels d'auteurs, oeuvres, noms de personnes, noms de lieux, matières avec des traits liés à la numismatique : 12 mois 6 mois IGE	VA_17_UCAEN	-

Equipe	Intitulé	Référence	Remarque
IRHT	Développement de référentiels diplomatique et apport de data : 6 mois et 24 mois	VA_18_CNRS	-
CIHAM	Développement de référentiels valeurs et mesures : 8 mois et 24 mois	VA_19_CNRS	-
HiSoMA	Alignement des métadonnées de Biblindex avec les référentiels Biblissima : 4 mois	VA_20_CNRS	-
HiSoMA	Alignement des données bibliques de Biblindex avec les référentiels : 2 mois	VA_21_CNRS	-
IRHT	Alignement pérenne et automatisé des BDD utilisant les référentiels de l'Orient chrétien : 18 mois	VA_22_CNRS	-
HiSoMA	Mécanisme d'automatisation des mises à jour de données Biblindex sur le portail Biblissima : 2 mois	VA_23_CNRS	-
CRH	Développement de l'interopérabilité des bases de données du CRH (images, exempla…) : 24 mois	VA_24_EHESS	-
GED	Accompagnement de la numérisation : 6 mois IE	VB_01_CC	-
CESR	Interopérabilité des données et des corpus textuels et image portail BVH → Biblissima et portails partenaires (Gallica, EDIT16, SUDOC, etc.) : 60 mois IGE	VB_02_UTOURS	-

Equipe	Intitulé	Référence	Remarque
AOROC	Réalisation de modèles numériques 3D et impression 3D : 12 mois IGE	VB_03_EPHE	-
IRHT	Projet CRMBF-3D : catalogue des reliures médiévales des bibliothèques de France – Inventaire, prise de vue 3D pour env. 5000 volumes : 18 mois	VB_04_CNRS	-
IRHT	Projet CRMBF-3D : catalogue des reliures médiévales des bibliothèques de France – Mise en ligne et publication des notices dans Bibale : 3 mois IE	VB_05_CNRS	-
Bbma	Accompagnement du déploiement de IIIF dans le réseau des archives dans le cadre de IIIF360 : 48 mois IGE	VB_06_CC	-
CRC	Mise en place des outils informatiques pour l'exploitation des données analytiques : 48 mois IR	VB_07_CNRS	-
IRHT	Classification des éléments graphiques (pages et zones de pages) : 6 mois IE	VB_08_CNRS	-
IRHT	Catalogage automatique des manuscrits numérisés : identification des textes issus de HTR par comparaison avec référentiels textuels (Corpus corporum etc.) : 40 mois IE	VB_09_CNRS	-
IRHT	Reconnaissance d'entités nommées (cote, noms de personnes, titres d'œuvres) et	VB_10_CNRS	-

Equipe	Intitulé	Référence	Remarque
	alignement sur des référentiels : 16 mois IR et 12 mois post doc		
AOROC	Développement de Kraken : 32 mois IR	VB_11_EPHE	-
IRHT	Répertoire de filigranes : création de métadonnées, missions photographiques : 6 mois IE	VB_12_CNRS	-
IRHT	Développement d'Extractor : 3 mois IR	VB_13_CNRS	-
CJM	Alimenter la base filigranes en métadonnées, en images dans les mss et les imprimés, également par science participative : 12 mois IGE	VB_14_ENC	-
AOROC	Acquisition des données numismatiques en musée et réserves et mise en ligne : 6 mois IGE	VB_15_EPHE	-
AOROC / Mines Paris	Réalisation d'un système automatique de reconnaissance des coins monétaires antiques : 4 mois AI	VB_16_EPHE	-
SAPRAT	Module héraldique - Suivi du projet et appariement des données : 33 mois IR	VB_17_EPHE	-
SAPRAT	Module Sigiscript (épigraphie du sceau et reconnaissance automatique) - Suivi du projet et appariement des données : 33 mois IR	VB_18_EPHE	-
AOROC	Développement/intégration d'Archetype : 12 mois IR	VB_19_EPHE	-

Equipe	Intitulé	Référence	Remarque
AOROC	Développement d'eScriptorium (temps plein) : 28 mois IR	VB_20_EPHE	-
AOROC	Maintenance d'eScriptorium : 24 mois IR	VB_21_EPHE	-
SAPRAT	Multipal - Développement, alimentation pour les systèmes graphiques qui ne sont pas encore traités : 22 mois IR	VB_22_EPHE	-
AOROC	Géoréférencement et spatialisation des données épigraphiques sur Chronocarto : 12 mois IE	VB_23_EPHE	-
AOROC	Intégration des données sur la base en ligne : 8 mois AI	VB_24_EPHE	-
HiSoMA	Production de 2 cours d'auto-formation à l'encodage et à la publication de sources TEI dans le modèle EpiDoc (épigraphie) : 6 mois IE	VB_25_CNRS	-
CESCM	Agrégation de nouveaux bassins de données : les inscriptions médiévales (CIFM, RICG, Royaume de Jérusalem), bibliographie et archives : 60 mois IR	VB_26_CNRS	-
CESCM	Agrégation de nouveaux bassins de données (développement TITULUS, édition XML-TEI) : 1 stage par an pendant 5 ans	VB_27_CNRS	-
CJM	CLEM Carmina Latina Epigraphica Moderna : 6	VB_28_ENC	-

Equipe	Intitulé	Référence	Remarque
	mois IGE		
CIHAM	Edition TEI des gloses. Développement informatique (environnements de balisage et outil de publication web) : 6 mois IE	VB_29_ENSLYON	-
IRHT	Expertises typologiques diverses sur les textes latins antiques et médiévaux. Production de données et de documentation érudite sur ces textes, dont édition, transcription et critique de textes. Référentiels d'autorités pour le monde latin. 40 mois IE + 5 mois IR + 12 mois post doc	VB_30_CNRS	-
HiSoMA	Liaison par API des données Biblindex avec d'autres projets internationaux – Suivi : 3 mois IR	VB_32_CNRS	-
HiSoMA	Développement d'une plateforme collaborative d'enrichissement des données de Biblindex – Suivi : 3 mois IR	VB_33_CNRS	-
HiSoMA	Développement d'interfaces de visualisation spatio-temporelle des données statistiques de Biblindex, réutilisables pour d'autres projets – Suivi : 12 mois IR	VB_34_CNRS	-
CJM	Chaînage des outils d'édition et d'étude des documents d'archives (cartulaires et chartriers) (B-I.6.2 et B-I.8) : annotation des données textuelles (repérage des entités nommées,	VB_35_ENC	-

Equipe	Intitulé	Référence	Remarque
	alimentation des référentiels) : 10 mois de vacations		
CJM	Chaînage des outils d'édition et d'étude des documents d'archives (cartulaires et chartriers) (B-I.6.2 et B-I.8) : : constitution de la chaîne éditoriale : 8 mois IGE	VB_36_ENC	-
CJM	Chaînage des outils d'édition et d'étude des documents d'archives (cartulaires et chartriers) (B-I.6.2 et B-I.8) : récolement et préparation d'un échantillon-test de documents numérisés : 16 mois de vacations	VB_37_ENC	-
CJM	Corpus numérique du droit romain (Miroir des classiques) : édition TEI des textes de droit romain en français : 12 mois post doc	VB_38_ENC	-
CJM	Automatiser la collation des textes xml et garder en sortie la structuration xml-TEI – collation assistée : 12 mois post doc	VB_39_ENC	-
CJM	Automatiser la collation des textes xml et garder en sortie la structuration xml-TEI – API, interface : 12 mois IGE	VB_40_ENC	-
IRHT	Développement de TELMA/ANACLET (en configuration CMS) et alimentation IRHT – dévelopement et préparation de corpus : 6 mois IR + 12 mois post doc	VB_41_CNRS	-

Equipe	Intitulé	Référence	Remarque
MRSH	Spécification et conception du laboratoire, animation scientifique de l'observatoire, mise en place d'un espace d'échange de schémas documentés : 44 mois IR	VB_42_UCAEN	-
MRSH	Développement et adaptation du moteur MaX, plugin de travail collaboratif, connecteurs pour ces outils (oXygen, DTS) : 20 mois IE	VB_43_UCAEN	-
MRSH	En phase d'exploitation de Biblissima+, animation scientifique du laboratoire, formations TEI, développements au fil de l'eau (emplois pérennisés) : 8 mois IGR et 4 mois IGE	VB_44_UCAEN	-
CRAHAM	Conception d'environnements adaptés aux différents types de sources anciennes, médiévales, Renaissance éditées par le CRAHAM et outillage de ces sources : 36 mois IGR	VB_45_UCAEN	-
CRAHAM	Editions et annotations de sources : 10 mois IGE	VB_46_UCAEN	-
CRAHAM	Tests sur les sources encodées en XML-TEI et réflexion avec le PDN et les autres partenaires sur l'outillage des sources : 8 mois IGE	VB_47_UCAEN	-
HiSoMA	En lien avec les réflexions menées au sein de l'observatoire, définition d'un protocole XML-TEI	VB_48_CNRS	-

Equipe	Intitulé	Référence	Remarque
	d'encodage des citations de la bible, expérimenté sur des échantillons variés de corpus et sur Biblindex : 6 mois IE		
IRHT	Configuration de pluCo pour Oxygen : 2 mois IR	VB_49_CNRS	-
IRHT	Développement de configurations types pour le moteur d'affichage Max : 2 mois IR	VB_50_CNRS	-
IRHT	Développement d'une solution conviviale pour le travail collaboratif dans Oxygen : suivi des développements de pluCo dans ce domaine et adaptation pour Oxygen : 2 mois IR	VB_51_CNRS	-
IRHT	Conception, réalisation, suivi des projets éditoriaux / Réflexion commune avec PDN et ENC sur la mutualisation des schémas d'encodage / Suivi de l'alignement des thesaurus d'autorités / Test des solutions de mise en ligne Max et teiPublisher / Développement des environnements de balisage sous Oxygen et méthodologies d'encodage : 11 mois IR	VB_52_CNRS	-
CIHAM / HiSoMA	Production de 2 cours d'auto-formation à l'encodage et à la publication de sources TEI : 6 mois IR CIHAM	VB_53_CNRS	-
CIHAM / HiSoMA	Production de 2 cours d'auto-formation à	VB_54_CNRS	-

Equipe	Intitulé	Référence	Remarque
	l'encodage et à la publication de sources TEI : 6 mois IE CIHAM		
CIHAM	Sessions annuelles d'accompagnement des cours d'auto-formation (exercices corrigés, question / réponses, interactions) : 1 mois IR	VB_55_CNRS	-
HiSoMA	Organisation de sessions annuelles d'accompagnement des cours d'auto-formation (exercices corrigés, question / réponses, interactions) : 1 mois IE en lien avec le CIHAM	VB_56_CNRS	-
MRSH	Organisation d'une école d'été pour la diffusion des outils et des méthodes mises en place. Soutien aux éditeurs scientifiques : 2 mois IR	VB_57_UCAEN	-
IRHT	Projet Relicantus : inventaire et numérisation de fragments musicaux – campagne de numérisation : 2 mois T	VB_58_CNRS	-
IRHT	Projet Relicantus : inventaire et indexation : 6 mois IE	VB_59_CNRS	-
IRHT	Projet Wala : indexation des sources musicales de l'ouest de la France : 12 mois post doc	VB_60_CNRS	-
CJM	Développement DTS, accompagnement, API : 43 mois IGR	VB_61_ENC	-

Equipe	Intitulé	Référence	Remarque
HiSoMA	Préparation de corpus de textes patristiques (grec, latin, syriaque) pour utilisation du protocole DTS : 6 mois IR	VB_62_CNRS	-
HiSoMA	Intégration des données et fonctionnalités du lemmatiseur Hisoma dans Eulexis : 1 mois IR	VB_63_CNRS	-
HiSoMA	Intégration de Collatinus et Eulexis dans la chaîne de traitement des données textuelles de Biblindex : 1 mois IR	VB_64_CNRS	-
IRHT	Corpus lexical européen (50 M mots de latin médiéval européen 700-1300) : 11 mois IR	VB_65_CNRS	-
HiSoMA	Mise en oeuvre d'outils stylométriques et textométriques de repérage d'intertextualité sur des textes latins médiévaux : 12 mois post doc	VB_66_CNRS	-
HiSoMA	Développement d'un outil générique de repérage de l'intertextualité : 6 mois IR	VB_67_CNRS	-
CJM	Centre de ressources computationnelles pour les langues à variation graphique : 42 mois IGR	VB_68_ENC	-
GED	Développeurs graphistes pour accompagnement de la médiation scientifique : 12 mois IGR	VB_69_CC	-

31 Projets Lauréats de l'appel à manifestation d'intérêt (AMI) – périmètre P3

Voir aussi la page https://projet.biblissima.fr/fr/appels-projets/projets-retenus/

Année	Titre projet	Etablissement porteur	Equipe associée	Catégorie de soutien	Responsable scientifique
2022	Bibliotheca Carnotensis Nova	Médiathèque l'Apostrophe - Ville de Chartres	IRHT - Institut de recherche et d'histoire des textes	Projet partenarial	Claudia Rabel ; Joanna Frońska
2022	Éditions critiques relatives à l'Université de Paris (ECRU)	Bibliothèque interuniversitaire de la Sorbonne (BIS)	IRHT - Institut de recherche et d'histoire des textes	Projet partenarial	Laurence Bobis ; Thierry Kouamé
2022	Fabliaux	CIHAM - Histoire, Archéologie, Littératures des mondes chrétiens et musulmans médiévaux	Institut d'Histoire des Représentations et des Idées dans les Modernités (IHRIM)	Projet partenarial	Corinne Pierreville
2022	Manuscrits Génovéfains (MAGE)	Bibliothèque Sainte-Geneviève (BSG)	IRHT - Institut de recherche et d'histoire des textes	Projet partenarial	Nathalie Rollet-Bricklin
2022	RESCAPÉ	Biblioteca Nazionale Universitaria di Torino	CJM - Centre Jean Mabillon	Projet partenarial	Marco Maulu
2022	Reverse Engineering Kennicott (REK)	AOrOc – Archéologie et philologie d'Orient et d'Occident	National Library of Israel	Projet partenarial	Daniel Stökl Ben Ezra
2022	Smart Critical Ben Sira (SCRIBES)	ÉCRITURES (EA 3943) - Université de Lorraine	CRAHAM - Centre Michel de Boüard	Projet partenarial	Jean-Sébastien Rey
2022	Venezia Libro Aperto (VeLA)	Università Ca' Foscari di Venezia, Dipartimento di Studi Umanistici	CESCM – Centre d'études supérieures de civilisation médiévale	Projet partenarial	Flavia De Rubeis

Année	Titre projet	Etablissement porteur	Equipe associée	Catégorie de soutien	Responsable scientifique
2023	Du papyrus hiératique au texte numérique	HiSoMA – Histoire et Sources des Mondes Antiques	-	Bourse Jeune Chercheur	Quentin Cécillon
2023	Un catalogue des travaux savants imprimés et manuscrits consacrés à Euripide au XVIe siècle	HiSoMA – Histoire et Sources des Mondes Antiques	Litt&Arts, Université Grenoble-Alpes	Bourse Jeune Chercheur	Alexia Dedieu
2023	EpiHorrea	CCJ - Centre Camille Jullian	AOrOc – Archéologie et philologie d'Orient et d'Occident	Projet exploratoire	Stéphanie Satre ;
					Bruno Baudoin
2023	HTRogène	CJM - Centre Jean Mabillon	CIHAM - Histoire, Archéologie, Littératures des mondes chrétiens et musulmans médiévaux	Projet exploratoire	Thibault Clérice
2023	Manuscrits des classiques latins de la bibliothèque Farnèse	IRHT - Institut de recherche et d'histoire des textes	MRSH-PDN – Pôle Document numérique ; Biblioteca Nazionale di Napoli ; Ecole française de Rome	Projet exploratoire	Angela Cossu
2023	Burgundia Scripta Merovingica	ARTEHIS	IRHT - Institut de recherche et d'histoire des textes ; Bibliothèque nationale de France (BnF) ;	Projet partenarial	Dominique Barbet-Massin ; Aurélia Bully ; Marie-José Gasse-Grandjean

Année	Titre projet	Etablissement porteur	Equipe associée	Catégorie de soutien	Responsable scientifique
			Bibliothèque municipale de Lyon ; Bibliothèque Bussy-Rabutin d'Autun ; Earlier Latin Manuscripts (base de données du Lowe, Codices Latini Antiquiores), Moore Institute		
2023	Hyper-Estampages	Ecole française d'Athènes	HiSoMA – Histoire et Sources des Mondes Antiques ; Bibliothèque de l'Institut de France	Projet partenarial	Michèle Brunet
2023	Les dessins de sceaux de la collection Gaignières	Ecole du Louvre	IRHT - Institut de recherche et d'histoire des textes ; SAPRAT-EPHE - Savoirs et pratiques du Moyen Âge à l'époque contemporaine ; Bibliothèque nationale de France (BnF)	Projet partenarial	Anne Ritz-Guilbert
2023	Les tablatures de la collection Albani	CESR - Centre d'études supérieures de la Renaissance	Ente Olivieri – Biblioteca e musei Oliveriani	Projet partenarial	Philippe Canguilhem
2023	SION Digit Sfarim	IRHT - Institut de recherche et d'histoire des textes	Archives départementales de la Gironde ; Archivio di Stato di Venezia	Projet partenarial	Evelien Chayes

V.III Fiches pratiques

32 Métadonnées d'un dépôt Zenodo

Cette fiche explicite les métadonnées minimales à utiliser pour le dépôt d'un jeu de données dans Zenodo[1].

Liens utiles :

- Accès : https://zenodo.org
- Bac à sable : https://sandbox.zenodo.org/
- Guide utilisateur : https://help.zenodo.org/docs/
- Tutoriels :
  - https://www.datacc.org/wp-content/uploads/2020/02/tutorielzenodov2.pdf
  - https://openscience.pasteur.fr/2022/12/07/comment-deposer-des-donnees-de-recherche-dans-zenodo/

Section	Champ Zenodo	Obligatoire	Définition	Recommandations / Exemples
-	Select a Community	Non	Choisir une ou plusieurs Communauté(s) Zenodo à laquelle associer le jeu de données	Pour un dépôt associé à un cluster de Biblissima+, sélectionnez Biblissima-cluster-N » (en remplaçant « N » par le numéro du cluster correspondant). Pour un dépôt associé au projet global, sélectionner « Biblissima ». – Il est possible de choisir plusieurs communautés. Biblissima+ recommande d'associer systématiquement le dépôt à d'autres communautés Zenodo institutionnelles (laboratoire, institution de tutelle, autre financeur etc.) si elles existent.
Basic information	Drag and Drop files or Upload files	Oui	Téléverser les fichiers associés au dépôt	Un dépôt doit contenir au moins un fichier numérique. Le volume total d'un dépôt est limité à 50 Go par défaut mais il est toujours possible de contacter les administrateurs de Zenodo en cas de volumes plus importants.
Basic information	Digital Object Identifier	Oui	Identifiant pérenne du dépôt	Un DOI (Digital Object Identifier) est un identifiant international stable et

Section	Champ Zenodo	Obligatoire	Définition	Recommandations / Exemples
				pérenne dont l'attribution est gérée par le consortium DataCite. Cliquez sur « No » pour laisser Zenodo attribuer automatiquement le DOI (il ne pourra pas être modifié ultérieurement) ou bien conservez la valeur « Yes » cochée, et ajoutez le DOI attribué par un éditeur ou un laboratoire. Attention ! le bouton "Get a DOI now!" vous permet de réserver un DOI depuis le formulaire afin de l'intégrer dans les fichiers versés dans le dépôt.
Basic information	Resource type	Oui	Type de ressource	Sélectionner un type dans la liste déroulante, par exemple « Dataset », « Software », « Model », « Presentation », etc.
Basic information	Title	Oui	Titre principal du jeu de données	Texte libre.
Basic information	Publication date	Oui	Date à laquelle le jeu de données a été publié	Par ex. « 2022-03-31 ».
Basic information	Creators	Oui	Personne(s) responsables(s) de la création du jeu de données	« Nom de famille, Prénom ». – Renseigner l'affiliation et si possible l'identifiant pérenne chercheur ORCID. Il est possible de préciser

Section	Champ Zenodo	Obligatoire	Définition	Recommandations / Exemples
				le rôle de l'individu ou de l'organisation dans la création de la ressource (par ex. « Data collector », « Data curator », « Host institution », « Contact person », etc.).
Basic information	Description	Non	Description sommaire	Texte libre. Indiquer la référence au livrable pour faciliter le suivi.
Basic information	Licence	Non	Licence du jeu de données	Sélectionner parmi les valeurs suggérées par Zenodo.
				Champ obligatoire si « Open Access » ou « Embargoed Access » ont été choisis dans le champ « Access right ».
Recommended information	Contributors	Non	Contributeurs	Idem champ « Creators ».
Recommended information	Keywords and subjects	Non	Mots clés ou indexation sujet dans un vocabulaire contrôlé.	Mots-clés libres. – Biblissima+ recommande l'utilisation des vocabulaires contrôlés utilisés par le moteur de recherche Isidore (cf. https://isidore.science/vocabularies).
Recommended information	Language	Non	Langue principale	Sélectionnez parmi les valeurs suggérées par Zenodo, ou à défaut saisir le code ISO 639 de la langue (sur deux

Section	Champ Zenodo	Obligatoire	Définition	Recommandations / Exemples
				ou trois lettres). – Par ex. « fr, en, ou Latin ».
Recommended information	Dates	Non	Dates clés	Indiquez au minimum une date. Plusieurs dates sont possibles. L'interface permet d'associer un type (par exemple : « Accepted », « Availabel », « Valid », etc.) et une description à chaque date ajoutée.
Recommended information	Version	Non	Numéro de version	Principalement utile pour les logiciels et les codes sources. L'utilisation d'un patron sémantique est recommandé (par exemple : MAJOR.MINOR.PATCH, ex. « 1.0.2 » – voir https://semver.org/
Recommended information	Publisher	Non	Personne physique ou morale responsable pour la publication du dépôt	Indiquez ici le nom de la personne physique ou morale devant figurer dans la citation.
Funding	Award	Non	Sources de financement	Cliquez sur le « Add award » et taper « Biblissima » dans le champ de recherche pour sélectionner l'EquipEx Biblissima+ (numéro : ANR-21-ESRE-0005).
Alternate identifiers	Identifier	Non	Identifiant supplémentaire	Indiquez les autres identifiants de ce dépôt.

Section	Champ Zenodo	Obligatoire	Définition	Recommandations / Exemples
Related works	Related works	Non	Travaux associés	Indiquez ici les identifiants de travaux en relation avec le dépôt. Préciser le type d'identifiant dans le champ « Scheme » (DOI, Handle, ARK, PURL, ISSN, ISBN, URn, Urls, etc.) et précisez le type de ressource.
References	References	Non	Référence associée	Texte libre.
Publishing information	Journal, Imprint, Thesis…	Non	Informations bibliographiques	Information bibliographiques d'une publication. Non pertinent pour les types jeu de données ou code source.
Conference	Title, place, website…	Non	Informations bibliographiques	Information bibliographiques d'une communication dans un congrès. Non pertinent pour les types jeu de données ou code source.
Draft	Save draft / preview / publish	Oui	Statut de publication du dépôt	Quand tous les champs obligatoires sont renseignés, vous pouvez publier le dépôt ou le soumettre aux modérateurs de la communauté choisie. Une prévisualisation est possible.
Visibility	Public / restricted / Apply an embargo	Oui	Accès public, avec embargo ou restreint.	Il est possible de restreindre l'accès aux fichiers déposés, soit de manière temporaire avec un embargo, soit

Section	Champ Zenodo	Obligatoire	Définition	Recommandations / Exemples
				de manière permanente. Les métadonnées quant à elles sont toujours libres d'accès.

1. Les métadonnées d'un dépôt sur Zenodo sont conformes au schéma de métadonnées Datacite cf. https://schema.datacite.org/ mais peuvent être exportés dans différents formats (Dublin Core, MARCXML, JSON-LD, etc.) ↩