Biblissima+ présente son projet pour la décennie à venir

Biblissima+ : Observatoire des cultures écrites de l'argile à l'imprimé

Douze unités de recherche en histoire, littérature, philologie, archéologie, sciences des matériaux et humanités numériques et computationnelles, le Grand Équipement Documentaire du Campus Condorcet, le Service Interministériel des Archives de France, les Archives nationales et l'entreprise TEKLIA s'associent pour construire Biblissima+, Observatoire des cultures écrites, de l'argile à l'imprimé. Ils ont rédigé ensemble un « livre blanc » qui présente en détail l'infrastructure numérique qu'ils souhaitent réaliser pour permettre de nouvelles découvertes.

AOROC – Archéologie et philologie d'Orient et d'Occident
UMR 8546 (CNRS – EPHE-PSL – ENS-PSL)
http://www.archeo.ens.psl.eu

CESCM – Centre d'études supérieures de civilisation médiévale
UMR 7302 (Université de Poitiers – CNRS)
https://cescm.labo.univ-poitiers.fr

CESR – Centre d'études supérieures de la Renaissance
UMR 7323 (Université de Tours – CNRS – Ministère de la Culture)
https://cesr.cnrs.fr

CIHAM – Histoire, Archéologie, Littératures des mondes chrétiens et musulmans médiévaux
UMR 5648 (CNRS – EHESS – ENS de Lyon – Université Jean Moulin Lyon 3 – Université Lumière Lyon 2 – Avignon Université)
http://ciham.msh-lse.fr

CJM – Centre Jean Mabillon
EA 3624 (ENC-PSL)
http://www.chartes.psl.eu/fr/rubrique-centre-jean-mabillon/centre-jean-mabillon

CRAHAM – Centre Michel de Boüard - CRAHAM
Centre de recherches archéologiques et historiques anciennes et médiévales
UMR 6273 (Université de Caen Normandie – CNRS)
http://www.unicaen.fr/craham/

CRC – Centre de Recherche sur la Conservation
USR 3224 (Muséum national d'histoire naturelle – CNRS – Ministère de la Culture)
http://crc.mnhn.fr

CRH – Centre de Recherches Historiques
UMR 8558 (CNRS – EHESS)
http://crh.ehess.fr

GED – Grand équipement documentaire du Campus Condorcet
https://www.campus-condorcet.fr/pour-la-recherche/grand-equipement-documentaire/une-bibliotheque-pour-la-recherche

HiSoMA – Histoire et Sources des Mondes Antiques
UMR 5189 (CNRS – ENS de Lyon – Université Jean Moulin Lyon 3 – Université Lumière Lyon 2 – Université Jean-Monnet Saint-Étienne)
https://www.HiSoMA .mom.fr

IRHT – Institut de recherche et histoire des textes
UPR 841 (CNRS)
https://www.irht.cnrs.fr

MRSH-PDN – Pôle Document numérique, MRSH
USR 3486 (Université de Caen Normandie – CNRS)
http://www.unicaen.fr/recherche/mrsh/document_numerique

SIAF – FranceArchives
https://francearchives.fr/fr/

AN – Archives nationales
http://www.archives-nationales.culture.gouv.fr/fr_FR/web/guest/home

SAPRAT – Savoirs et pratiques du Moyen Âge au XIX^{e} siècle
EA 4116 (EPHE-PSL)
http://saprat.ephe.sorbonne.fr

TEKLIA - SAS
https://teklia.com

Biblissima+ : Observatoire des cultures écrites de l'argile à l'imprimé

Biblissima+ est 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, il y a 3 000 ans, aux premiers livres imprimés, quels que soient les supports et les langues. Biblissima+ met l'accent au départ sur les collections patrimoniales conservées en France par les bibliothèques, les services d'archives, les musées.

Biblissima+ crée un portail national réalisant **l'interopérabilité de ressources numériques hétérogènes totalement open source**, et donnant ainsi un accès facilité à la documentation écrite originale, à la bibliographie et aux archives de la recherche. Il y ajoute un **environnement de travail exceptionnel proposant des chaînes d'outils numériques de pointe** pour enrichir, partager, réutiliser les corpus selon les principes FAIR.

Biblissima+ fédère, sur l'ensemble du territoire national, **15 équipes de recherche travaillant sur les textes, de l'Antiquité à l'édition numérique**, dont une micro-entreprise et le Service Interministériel des Archives de France, qui n'avaient jamais travaillé toutes ensemble de façon concertée, mais dont 8 étaient déjà impliquées dans l'équipex Biblissima (2012-2020). Les équipes partenaires, qui développent les briques principales de l'infrastructure (ressources scientifiques et outils innovants), forment autant de **pôles de compétence à l'échelle nationale**, en région (Aubervilliers, Pierrefitte, Caen, Lyon-Avignon, Orléans, Poitiers, Tours) ou à Paris. Cette structuration leur permet de concilier politique de site (en particulier en lien avec les CPER et les DRAC) et politique nationale et internationale.

Un système d'appels à projets ouvert à tous, destiné à produire de nouveaux jeux de données interopérables et de nouveaux outils, permet **d'intégrer progressivement de nouvelles données et de nouvelles communautés**. Un volet de l'appel à projets est international. Le portail Biblissima+ accentue cette ouverture internationale, à travers son moteur de recherche IIIF Collections, conçu pour interroger ensemble toutes les bibliothèques numériques interopérables du monde.

Biblissima+ se construit en 5 ans, et consacre les 3 dernières années à la mise en interopérabilité de tous les jeux de données produits, à la mise à jour constante de l'environnement de traitement des données et à l'enrichissement de la chaîne d'outils via les projets partenariaux. Le Campus Condorcet et l'Université de Caen pérennisent pour cela 4,5 emplois (équipe Biblissima+ et Pôle Document Numérique).

A. **Le premier volet décrit les mécanismes d'agrégation de nouveaux bassins de données dans le portail Biblissima+, via les référentiels Biblissima+.**

Biblissima+ est à l'échelle nationale le lieu de traitement et de mise en interopérabilité de toutes les données en jeu dans l'histoire de la transmission et de l'étude des cultures écrites anciennes, en cohérence avec la création sur le Campus Condorcet d'un Institut des langues rares (ILARA), soutenue en 2020 par le MESRI. Biblissima+ concerne 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.

L'agrégation des données se fait autour d'un **enrichissement massif des référentiels data.biblissima**, qui seront mis à la disposition du FNE (Fichier national d'entités) et des grands acteurs du monde de la conservation et de la diffusion scientifique, afin de leur permettre d'affiner la fouille de données dans la documentation qu'ils mettent à disposition des usagers. **En retour, Biblissima+ agrège autour de ses identifiants la numérisation des sources anciennes**, **les ressources documentaires** (catalogues en ligne, bases de données, archives de la recherche numérisées), **la bibliographie** (éditions électroniques, articles et ouvrages archivés, publications vivantes). On verra s'articuler et s'enrichir mutuellement différents outils nationaux de grande envergure sur une thématique large et variée, qui se trouve à la base de toute histoire de la culture. Ainsi d'importants bassins de données deviendront interopérables, comme le souhaitent les utilisateurs.

Second grand mécanisme d'enrichissement : **l'automatisation de la mise à jour des données mises en interopérabilité profonde** par l'infrastructure numérique, afin de garantir dans le temps l'efficacité de celle-ci.

B. Le second volet est centré sur l'utilisateur et concerne le portail de consultation et l'environnement d'enrichissement des données fournissant des outils de traitement innovants.

Les briques proposées par les pôles de compétence ont permis d'identifier **7 grands domaines à la pointe de l'innovation**, qui forment autant de clusters regroupant chercheurs, conservateurs et ingénieurs. **Les 7 clusters sont organisés selon le cycle de vie des données** :

1/ Acquisition des corpus de sources interopérables
2/ Prise en compte et cherchabilité des données d'analyse des matériaux
3/ Intelligence artificielle, reconnaissance de formes et d'écritures manuscrites
4/ Traitement approfondi des systèmes graphiques et analyse des documents
5/ Edition de sources en TEI
6/ Défis du patrimoine musical
7/ Interopérabilité et analyse des textes

Chaque cluster produit des outils, a sa page et son entrepôt de données sur le site de Biblissima. Chacun a un coordinateur qui interagit constamment avec l'équipe coordinatrice Biblissima+. L'équipe Biblissima+ travaille avec eux au dialogue entre clusters et au chaînage des outils, au plus près des utilisateurs, et réalise l'interopérabilité des résultats. Ces clusters structurent ainsi les communautés. Ils produisent, organisent et mutualisent les outils, réduisant les redondances, favorisant les inventions et leur réutilisation.

Biblissima+ est ainsi **une infrastructure et un écosystème capables de s'enrichir perpétuellement**. Un **portail Biblissima+ simplifié** permet de constituer des jeux de données, et de **rebondir vers la galaxie des outils**.

Biblissima+ relève ainsi plusieurs défis :

☞ **Structurels et sociologiques** : Biblissima+ fait travailler ensemble les communautés de littéraires, de philosophes, d'historiens, d'archéologues, d'historiens de l'art, d'épigraphistes, de numismates, de conservateurs de bibliothèques, archives et musées ayant affaire aux documents portant du texte, sans distinction d'aire culturelle ou d'époque. L'infrastructure les réunit pour la première fois de façon cohérente, leur permettant de dépasser leurs différences, de dialoguer, d'échanger leurs compétences, et de les accroître de façon significative, dans une logique d'écosystème. Elle investit dans l'ergonomie et la facilitation pour amener l'ensemble des communautés à modifier leurs usages et progresser significativement dans la maîtrise des outils numériques les plus innovants.

☞ **Techniques** : Biblissima+ accompagne ces communautés de haute érudition dans la conversion numérique et computationnelle, en particulier dans l'appropriation de l'intelligence artificielle, mais aussi dans le dialogue fécond avec la physique et la chimie, redéfinissant ainsi les humanités du 21e siècle. L'infrastructure aide de même les communautés de sciences dures à s'emparer des défis que leur offrent les humanités, avec leurs données complexes, leur culture de l'exception, leurs notions floues. L'infrastructure accélère enfin l'acculturation du monde de la conservation aux formats numériques et interopérables, qui aujourd'hui contribuent à la conservation des artefacts anciens et à leur exposition à tous les publics.

☞ **Scientifiques** : Biblissima+ a pour but d'accélérer cette évolution des communautés SHS travaillant sur la mémoire écrite de l'humanité pour leur permettre d'accéder à des informations que personne ne domine aujourd'hui. On en attend un changement de paradigme dans la découverte des textes de toutes les cultures, leur contextualisation, leur compréhension, leur mise à disposition, leur intégration dans le discours scientifique et dans la mémoire collective.

☞ **Sociétaux** : le tournant computationnel des sciences des textes ne concerne pas que la recherche, la conservation et la formation, mais tout citoyen, de tout âge, de tout milieu, de toute culture, dont les documents anciens sont le patrimoine, et dont la mémoire et l'identité se nourrissent de l'écrit. L'infrastructure développe les technologies les plus sophistiquées de la recherche fondamentale pour faciliter la mise à disposition de tous des sources et des résultats de la science des textes, par la transcription automatique, mais aussi par la traduction et l'audiovisuel.

Biblissima+ : Observatory of Written Cultures from Clay to Print

Biblissima+ is a multi-site digital infrastructure for research and service dedicated to the history of the transmission of ancient texts, from the first Mesopotamian clay tablets 3,000 years ago to the first printed books, in all media and languages. Biblissima+ initially focuses on heritage collections preserved in France by libraries, archives and museums.

Biblissima+ creates a national portal achieving the **interoperability of heterogeneous digital resources that are entirely open source,** thus providing easy access to original written documentation, bibliography and research archives. It adds an **exceptional working environment offering state-of-the-art pipeline of digital tools** to enrich, share and reuse corpora according to FAIR principles.

Biblissima+ federates **15 research teams throughout France working on texts, from Antiquity to digital publishing,** including a micro-enterprise and the Interministerial Service of the Archives of France, which had never before worked together in a concerted manner, but of which eight were already involved in the Biblissima Équipex (2012-2020). The partner teams develop the main building-blocks of the infrastructure (scholarly resources and innovative tools), and form **centres of expertise at the national level**, in the regions (Aubervilliers, Pierrefitte, Caen, Lyon-Avignon, Orléans, Poitiers, Tours) and in Paris. This structuring enables them to reconcile local policy (particularly in relation to the CPERs and DRACs) with national and international policy.

A system of open calls for proposals designed to produce new interoperable datasets and tools, allows for the **gradual integration of new data and communities**. One strand of the call for projects is international. The Biblissima+ portal accentuates this international openness through its IIIF Collections search engine, designed to access all the interoperable digital libraries of the world in a single query.

Biblissima+ is built in 5 years, and devotes the last 3 years to the interoperability of all the datasets it produces, to the constant updating of the data processing environment and to the enrichment of the pipeline of tools through partnership projects. Campus Condorcet and the University of Caen have created 4.5 new posts for this purpose (Biblissima+ team and Pôle Document Numérique).

A. **The first strand describes the mechanisms for aggregating new datasets in the Biblissima+ portal via the Biblissima+ authority files.**

Biblissima+ operates at the national level as the place for processing and interoperability of all the data at stake in the history of the transmission and study of ancient written cultures, in line with the creation on the Campus Condorcet of an Institute of Rare Languages (ILARA), supported in 2020 by MESRI. Biblissima+ concerns all heritage collections that preserve ancient texts, including archaeological sources, seals and coins, but also the archives of modern scholars and contemporary researchers.

The data aggregation is based on a **massive enrichment of the data.biblissima authority files**, which will be made available to the FNE (Fichier national d'entités) and the major players in the world of conservation and scholarly dissemination, to enable them to refine their data mining in the documentation they make available to users. **In return, Biblissima+ aggregates around its identifiers the digitization of ancient sources**, **documentary resources** (online catalogues, databases, digitized research archives), **bibliography** (electronic editions, archived articles and works, ongoing publications and series). Various large-scale national tools will be articulated and mutually enriched on a broad and varied theme, which is at the basis of any history of culture. In this way, important datasets will become interoperable, as desired by users.

The second major enrichment mechanism is **the automation of updating data that are deeply interoperable** by the digital infrastructure, in order to guarantee its efficiency over time.

B. **The second strand is user-centred and concerns the access portal and the data enrichment environment which provide innovative tools for processing.**

The bricks proposed by the competence clusters have made it possible to identify **7 major areas at the cutting edge of innovation,** which form groups of researchers, curators and engineers. The **7 clusters are organised according to the data life cycle** :

1/ Acquisition of interoperable source corpora
2/ Study and searchability of data issued from the analysis of materials
3/ Artificial intelligence, pattern recognition and handwriting recognition
4/ In-depth treatment of graphic systems and document analysis
5/ Editing sources in TEI
6/ Challenges of musical heritage
7/ Interoperability and text analysis

Each cluster produces tools and has its own page and data store on the Biblissima site. Each has a coordinator who constantly interacts with the Biblissima+ coordinating team. The Biblissima+ team achieves interoperability of results, working with the coordinators in a dialogue between clusters and the pipelines of tools while staying as close as possible to the users. These clusters thus structure the communities. They produce, organize and mutualise tools, reducing redundancies, promoting inventions and their reuse.

Biblissima+ is thus **an infrastructure and an ecosystem capable of perpetual enrichment**. A **simplified Biblissima+ portal** makes it possible to build up data sets, and to **bounce back to the galaxy of tools**.

Biblissima+ thus takes up several challenges:

☞ **Structural and sociological**: Biblissima+ brings together communities of literary scholars, philosophers, historians, archaeologists, art historians, epigraphers, numismatists, curators of libraries, archives and museums dealing with documents bearing text, without distinction of cultural area or period. The infrastructure brings them together for the first time in a coherent way, allowing them to overcome their differences, to dialogue, to exchange and significantly increase their skills, all in the logic of a coherent ecosystem. It invests in ergonomics and facilitation to bring all communities to enhance their uses of digital methods and to make significant progress in mastering the most innovative digital tools.

☞ **Technical**: Biblissima+ accompanies these communities of advanced scholarship in their conversion to the digital and computational, particularly in the appropriation of artificial intelligence, but also in fruitful dialogue with physics and chemistry, thereby redefining the humanities of the 21st century. The infrastructure also helps communities in the 'hard' sciences to take up the challenges offered by the humanities, with their complex data, their culture of exception, their fuzzy notions. Finally, the infrastructure accelerates the acculturation of the conservation world to digital and interoperable formats which today help preserve ancient artefacts and expose them to all audiences.

☞ **Scholars**: Biblissima+ aims to accelerate this evolution of SSH communities working on the written memory of humanity, in order to enable them to access information that noone masters today. It is expected to bring about a paradigm shift in the discovery of texts from all cultures, as well as in their contextualization, understanding, availability and integration into scholarly discourse and collective memory.

☞ **Societal**: the computational turn in the study of written culture does not only concern those in research, conservation and training, but every citizen, of any age, from any background, from any culture, whose ancient documents are their heritage, and whose memory and identity are nourished by the written word. The infrastructure develops the most sophisticated technologies of primary research to facilitate the availability the sources and results of study to all, through automatic transcription, but also through translation and audiovisual means.

… de Biblissima à Biblissima+ …

L'équipement d'excellence Biblissima, Observatoire du patrimoine écrit du Moyen Âge et de la Renaissance (équipex 0007), est né en 2012 grâce à une aide de 7,1 millions d'€ accordée par l'Etat au titre des Investissements d'avenir, qui a entraîné de la part des partenaires fondateurs un important investissement en données, en temps de travail et en environnement (pour un coût complet estimé initialement à 24 039 316 €), et la recherche active de cofinancements. Cet Observatoire est un « équipement de données » sur la circulation des textes en Occident du 8e au 18e siècle, principalement en latin, grec, hébreu et ancien français. Adapté aux besoins des SHS, de leurs structures et de leurs métiers (recherche, formation, conservation, médiation…), il est destiné à accumuler les données nécessaires pour l'exercice plein de ces métiers et à les rendre interopérables, afin de dépasser les remparts que constituent l'hétérogénéité des formats et des vocubulaires, en donnant accès à un certain nombre d'outils permettant de déployer le cycle de vie des données, de leur création à leur enrichissement puis leur réutilisation, en respectant et promouvant les principes FAIR (*Findable*, *Accessible*, *Interoperable*, *and Reusable*). Il contribue ainsi à la conversion numérique des humanités en France et dans le monde.

Biblissima, aujourd'hui très connu à l'international, est devenu une « marque » ; le projet international *MMM*, *Mapping Manuscript Migrations*, dont le partenaire français est l'IRHT, s'en est inspiré et partage avec lui l'une de ses quatre ressources, la base Bibale, mais Biblissima n'a pas encore d'équivalent. Sa réussite, soulignée en 2017 par le jury international lors de l'audit des équipex, et que confirme le nombre constamment croissant des visites (+ 81% en 2019 par rapport à 2018), tient à :

– la cohérence du projet, qui couvre tout le cycle de vie des données de la recherche, jusqu'à la construction de ses publics par la formation, la médiation, la promotion des formats ouverts et interopérables, l'ouverture totale des données, en passant par le sauvetage et la mise à disposition des données de projets terminés ou privés de financement et autrement voués à l'obsolescence (c'est le cas des données du projet européen *Europeana Regia*, dont le site est à l'arrêt depuis des années, et dont les données sont sauvées et toutes consultables et interopérables grâce au portail Biblissima) ;

– son périmètre scientifique vaste et en même temps compréhensible : l'histoire de la circulation des textes et des livres au Moyen Âge et à la Renaissance, essentiellement en Occident, à l'aide de toute la documentation des partenaires portant sur ces textes et ces objets, de l'Antiquité à nos jours ;

– la simplicité d'utilisation du portail et des outils, fondée sur des alignements sérieux et donc des référentiels robustes donnant accès à des données très complexes, et leur adéquation aux besoins ;

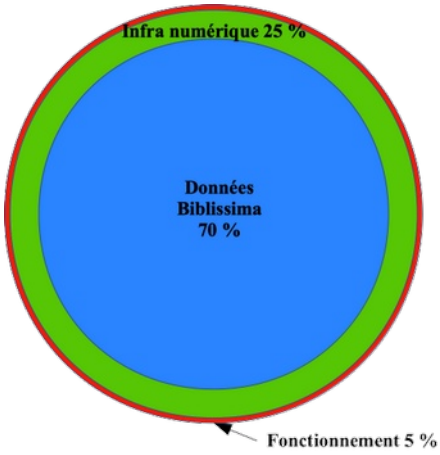
– sa capacité de fédérer et structurer des communautés différentes en les faisant travailler, chacune selon ses méthodes, à un but commun : le rapprochement des communautés de recherche et de formation d'une part, et de conservation et de médiation d'autre part est l'une des clés de son succès et de sa solidité, tant à l'échelle nationale qu'à l'échelle internationale.

L'équipex Biblissima représente en fait la phase initiale d'un processus qui doit prendre une ampleur tout autre, il **constitue les fondations d'une infrastructure nationale à vocation internationale, Biblissima+.**

A urgence égale, Biblissima avait une majeure et une mineure en termes de dépenses.

La majeure a été l'acquisition des données, car le but était de rassembler un corpus massif de données sur la transmission des textes anciens en Occident du 8e au 18e siècle, pour permettre de nouvelles recherches dans ce domaine, de nouveaux questionnements, des résultats inattendus. **Environ 70% de l'aide équipex a été consacrée à l'acquisition, la collecte, la mise en ligne de données**, en travaillant autant à la sauvegarde et à la diffusion numériques de données anciennes qu'à la production de données neuves, fondée sur l'exploration de gisements documentaires mal connus et l'usage de nouvelles technologies ; ont été financées des prestations et surtout des ressources humaines.

La mineure, en termes financiers, a été la création, le développement et l'enrichissement d'une infrastructure numérique rendant interopérables toutes ces données, et qui donne son sens au projet : **environ 25% de l'aide équipex a été consacrée au portail Biblissima**, porte d'entrée unique dans les ressources interopérables, et à l'ensemble des sites web donnant accès aux outils et à la documentation du projet. Les **5% restants** correspondent aux frais liés à la gouvernance et à la gestion du projet.



Aujourd'hui, le premier objectif est globalement réalisé : certains projets n'ont pu être achevés, mais d'autres, qui n'étaient pas prévus, ont été développés, de telle sorte que la masse critique de données prévue par le projet est atteinte, parfois dépassée, elle est archivée, et progressivement mise en interopérabilité par le portail (plus de 620 000 entités sont présentes à ce jour dans le portail [v. le détail infra, ill.], il y en aura un million fin 2020). Les partenaires continuent de développer ces ressources au fil de l'eau sur leurs fonds propres et en répondant à des appels à projets comme ils l'ont fait pendant toute la durée de l'équipex. La première vague de traitement et d'intégration des données dans le portail s'est focalisée sur l'histoire des collections anciennes de manuscrits et imprimés. La version courante du portail Biblissima intègre des jeux de données issus de 14 sources :

- **Esprit des livres** (Ecole nationale des chartes) - Catalogues de vente de bibliothèques de l'époque moderne (elec.enc.sorbonne.fr/catalogueVente/).
- **Europeana Regia** (BnF, BSB, BHUV, HAB, KBR) - Bibliothèque virtuelle portant sur les grandes collections royales de manuscrits en Europe au Moyen Age et à la Renaissance.
- **Bibale** (IRHT) - Base de données sur la transmission des textes médiévaux en Occident et l'étude des collections anciennes et modernes de manuscrits et livres imprimés anciens et de leurs possesseurs (bibale.irht.cnrs.fr).
- **CRII** (BVH, CESR - Université de Tours) - Catalogues régionaux des incunables informatisés.
- **Manuscripta Medica** (SAPRAT-EPHE, CIHAM) - Base de données de manuscrits médicaux latins (www.manuscripta-medica.com).
- **RegeCart** (IRHT) - Regestes de cartulaires (regecart.irht.cnrs.fr).
- **Bibliothèques françoises** (BVH, CESR - Université de Tours) - Edition numérique et outil de gestion de l'information des deux premiers dictionnaires imprimés d'auteurs français, celui de François Grudé, sieur de La Croix du Maine (Paris, 1584) et celui d'Antoine Du Verdier, sieur de Vauprivas (Lyon, 1585) (bibfr.bvh.univ-tours.fr)
- **BnF Archives et Manuscrits** - Bibliothèque de l'Arsenal (archivesetmanuscrits.bnf.fr)
- Manuscrits médiévaux de la Bibliothèque de **Wellcome Collection** (Wellcome Library, Londres) (wellcomelibrary.org)
- **Mandragore** (BnF) - Base des manuscrits enluminés de la Bibliothèque nationale de France (mandragore.bnf.fr)
- **Pinakes** (IRHT) - Base des manuscrits grecs (pinakes.irht.cnrs.fr)
- **BnF Archives et Manuscrits** - Département des manuscrits (archivesetmanuscrits.bnf.fr)
- **Comparatio** (IRHT) - Base de données sur les chants liturgiques médiévaux (IRHT-CNRS) (comparatio.irht.cnrs.fr)
- **Initiale** (IRHT) - Catalogue de manuscrits enluminés (initiale.irht.cnrs.fr)

Les entités présentes dans le portail Biblissima (juillet 2020)

La dernière source mise en intéropérabilité avec les précédentes au début du mois de juillet 2020, *Initiale*, permet aux chercheurs d'interroger ensemble, pour la première fois, l'intégralité des deux grandes bases de données sur les enluminures des manuscrits conservés en France, *Mandragore* (BnF) et *Initiale* (IRHT), leurs

descripteurs ayant été alignés autant qu'il était possible. Depuis septembre 2020, un **thesaurus iconographique de plus de 27 000 entrées** réconciliant les vocabulaires des deux bases facilite leur interrogation simultanée, donnant un accès thématique à leurs quelque 230 000 images (et plus de 306 000 notices).

Ces données ne deviennent un « équipement de données » et ne prennent tout leur sens que par la réalisation de l'infrastructure numérique et l'ensemble de ses services. Infrastructure et offre de services se sont déployés largement au-delà des attentes, et doivent encore se développer au rythme des évolutions technologiques et de l'évolution de la recherche en SHS, en grande partie due, dans le domaine de l'histoire de la culture écrite occidentale, au portail lui-même et aux dynamiques mondiales dans lesquelles il a joué un rôle au cours des 7 dernières années.

Nous sommes donc à un tournant, où le succès de Biblissima impose un nouveau départ et une inversion de la majeure et de la mineure : c'est le projet Biblissima+.

Désormais, même si l'exhaustivité est impossible et loin d'être atteinte, l'équipement de données prévu par le projet existe, il constitue un socle, et s'il faut continuer à produire et fédérer des données, cela ne doit être financé par le projet qu'à la marge, pour garantir l'innovation, une « respiration » du projet, rôle qu'ont joué très efficacement, dans Biblissima, les appels à projets annuels, qu'il faudrait poursuivre. **L'essentiel des accroissements doit venir désormais de la dynamique du portail lui-même**, capable d'agréger constamment de nouveaux contenus venus de ses partenaires mais aussi de l'extérieur de son partenariat initial, **sur le modèle de ce que fait déjà la plate-forme IIIF Collections créée par Biblissima**.

De ce fait, le portail et ses outils deviennent, dans le nouveau projet, la majeure. C'est de la qualité de sa conception, de son mécanisme d'accroissement, des nouveaux partenariats engagés, de l'ergonomie de son interface et de l'intérêt des corpus de données interopérables que vont naître véritablement l'infrastructure nationale Biblissima+ et son rayonnement international. A cet égard, **l'initiative *IIIF Collections of Manuscripts and Rare Books* de Biblissima sert de leçon** : IIIF Collections est un moteur de recherche fédérant progressivement les numérisations de manuscrits accessibles via les standards IIIF partout dans le monde, sans distinction de langue. Pour le moment, il donne un accès rapide à 73 000 documents (interrogation sur n'importe quelle entité : auteur, oeuvre, lieu…). IIIF Collections rencontre un grand succès international (84% des visites viennent de l'étranger, USA en tête) et a reçu de nombreuses manifestations d'intérêt (Allemagne, Pologne, Belgique, RU, Suisse, Italie), parce qu'il **répond à un besoin d'investigations larges dans un bassin de données très ouvert**. La prochaine refonte du portail Biblissima réunira les fonctions de IIIF Collections et du portail, afin d'augmenter significativement le nombre de documents auxquels il donne accès et que ses référentiels permettent d'agréger autour de tel ou tel objet (par ex. un texte, un auteur, une collection, un sujet iconographique etc.). La moisson de données sera simplement plus riche et approfondie pour les ressources des bibliothèques de France, mais il y aura aussi des surprises concernant les manuscrits conservés à l'étranger dans les bases de données françaises, et nous répondrons ainsi à l'exigence de **tenir compte de l'ensemble des bibliothèques du monde pour reconstituer les phases anciennes de l'histoire des cultures écrites**.

Solidement construite sur le socle de données, d'initiatives et de partenariats de l'équipex auquel elle doit son nom, **l'infrastructure Biblissima+ sera donc déployée sur un périmètre beaucoup plus large pour tenir compte des besoins de la recherche et des institutions qui la portent ou l'utilisent, et qui se sont fait jour au fil du développement du projet.** Ces besoins correspondent aux recommandations du jury international équipex (printemps 2017), qui souhaite que Biblissima transcende les frontières actuelles, qui n'existaient pas au Moyen Âge « *si bien qu'une pleine compréhension du flux des idées doit transcender l'hexagone* », et suggère de « *continuer d'élargir leur portée internationale et trouver des moyens de relier leurs ressources uniques à des documents similaires ailleurs* ». **Biblissima+ sera un Observatoire des cultures écrites anciennes**, centré comme auparavant sur le patrimoine écrit, avec un intérêt premier pour les textes anciens et leur transmission, mais

- sans limite de temps,
- sans limite de langue,
- sans limite de type de documentation (artefacts anciens de toute sorte, bibliographie ancienne et moderne, archives historiques ou scientifiques, documents sonores, simulations et vision par ordinateur etc.),
- avec toute la palette des sciences en interaction avec les disciplines plus traditionnelles dans le champ des humanités, qui aujourd'hui comprennent les sciences dures, les sciences cognitives, l'intelligence artificielle, aboutissant à une redéfinition large des disciplines de l'érudition,
- en accordant une attention toute particulière à son ouverture internationale.

Il en résultera une **infrastructure numérique multipolaire** reposant sur le réseau et le maillage national des pôles de compétences présentés en BI (pages 26 à 32), 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. L'accent sera mis sur les **collections de documents patrimoniaux conservées en France** par les bibliothèques, les services d'archives, les musées. Biblissima+ crée un portail donnant un accès facilité à la documentation écrite originale, à la bibliographie et aux archives de la recherche la concernant, ainsi qu'à un **environnement de travail proposant des chaînes d'outils pour enrichir, partager, réutiliser les corpus**. Son but est de renouveler en profondeur la recherche sur les cultures écrites anciennes en France et dans le monde, et de faire système à l'échelle du Campus Condorcet et à l'échelle **régionale** (en lien avec le CPER et le projet de centre en humanités numériques du Campus Condorcet, Condornum, et avec les CPER dans lesquels sont impliquées les équipes de Caen, Lyon, Poitiers, Tours, déposés fin 2019), **nationale** (par le rôle qu'y jouent les pôles de compétence régionaux, mais aussi en lien avec les IR et TGIR et les acteurs du patrimoine écrit) et **internationale** (en lien avec les infrastructures, comme DARIAH, et les consortia du domaine, comme IIIF ou le CERL [Consortium of European Research Libraries], ainsi que toutes les institutions concernées scientifiquement).

Biblissima+ permet de passer du développement d'un projet très riche à une intégration dans un écosystème d'outils interconnectés au service de la recherche. C'est une deuxième impulsion qui permet de faire passer l'infrastructure de l'équipement de données très enrichies et du portail performant à un dispositif enchâssé pour exploiter le patrimoine écrit, qu'il soit sous tutelle Culture ou ESR, et pour valoriser les outils développés par les communautés.

Le projet d'infrastructure qu'on va lire explicite donc d'abord **les mécanismes scientifiques et partenariaux d'élargissement du périmètre** de Biblissima aux niveaux national et international, ainsi que son rôle structurant dans le paysage français : **A. Un observatoire des cultures écrites anciennes**.

Il décrit ensuite **la façon dont l'infrastructure numérique multipolaire,** au carrefour des disciplines de l'érudition et des autres sciences, **peut jouer le même rôle fédérateur pour les outils** et **favoriser le développement de services** pour la recherche : **B. Un arsenal et un écosystème pour de nouvelles découvertes**.

A. Un Observatoire des cultures écrites anciennes

le périmètre scientifique et les mécanismes d'agrégation des données

L'existence d'un portail comme Biblissima, offrant la possibilité de fouiller dans une même interface une quantité de ressources hétérogènes, entraîne un désir d'exhaustivité, induit même une exigence d'exhaustivité. L'utilisateur averti ne comprend pas qu'il y manque des données dont il connaît l'existence par ailleurs, et dont il a besoin qu'elles soient rassemblées par le portail. L'utilisateur non averti manque une partie de l'information et ne sait pas nécessairement qu'elle existe, ni a fortiori où la chercher. Un tel portail peut donc reposer sur un choix de ressources de ses partenaires pour faire une preuve de concept à grande échelle – c'est ce qu'a fait Biblissima –, mais pour constituer une infrastructure nationale de recherche, il doit aller plus loin : il doit élargir le spectre des ressources, et il doit se projeter dans le temps, c'est-à-dire avoir un mode de fonctionnement permettant l'agrégation de nouvelles ressources – c'est déjà le cas –, et l'automatisation de leur mise à jour – ce n'est pas le cas, c'est difficile à faire, et c'est l'un des enjeux de la nouvelle infrastructure.

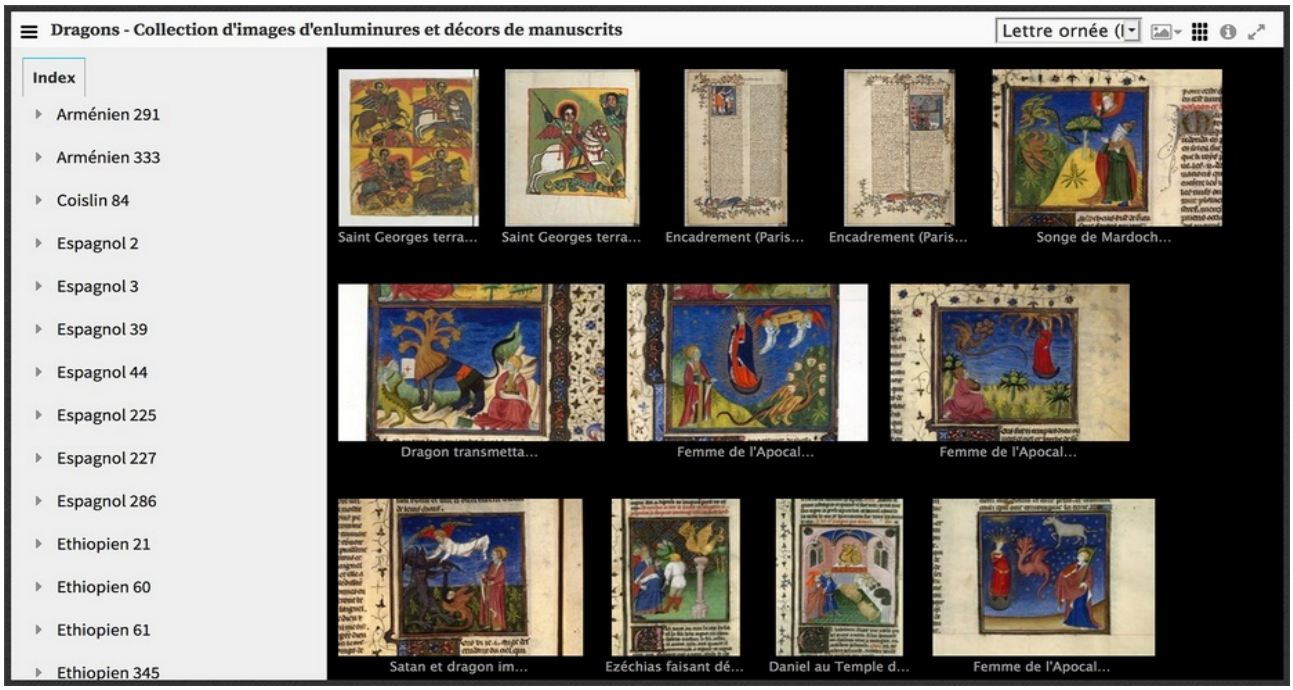
Cette partie présente donc les **logiques et mécanismes d'élargissement du bassin de données agrégées par le portail via ses référentiels** : données de signalement des documents, données bibliographiques, archives de la recherche, publications vivantes en open access. Biblissima+ souhaite mettre en place les conditions lui permettant de **s'interfacer au mieux avec les principaux programmes et outils dont s'est dotée la France, en montrant comment les articuler ensemble** : le CCFr, HAL, Collex Persée et Istex, OPERAS, en lien avec le GED et le nouvel Institut des langues rares sur le Campus Condorcet, et en lien avec les données des Archives nationales et de FranceArchives, portail national des archives de France que vient d'intégrer le GED.

A-I – VERS DE NOUVEAUX PARTENARIATS

A-I.1/ Toutes les cultures écrites anciennes et leurs vecteurs : articulation avec l'ILARA

Implication : tous les partenaires

Le premier Biblissima, en cohérence avec le programme scientifique qu'il s'était fixé, s'est limité aux grandes langues de culture de l'Occident médiéval et renaissant qui correspondaient aux spécialités de ses principaux partenaires : le grec et le latin, l'hébreu, l'ancien français. Ce périmètre linguistique a déjà été dépassé par la mise en interopérabilité sur le portail Biblissima de la base de données iconographiques de la BnF Mandragore, qui donne accès à tous les fonds manuscrits de la BnF. Une recherche quelconque sur un descripteur iconographique en démontre déjà l'intérêt dans une perspective comparatiste à l'échelle des cultures du monde (ill. : descripteur « dragons », qui montre une partie des réponses dans les fonds arménien, espagnol, éthiopien et Coislin [manuscrit transmettant des textes latins et grecs]). Dès lors que le premier programme scientifique est réalisé et les outils mis en place, il est en fait plus logique, pour les établissements de conservation comme pour les utilisateurs, de ne pas se limiter à certains fonds selon des critères linguistiques. **Biblissima+ s'ouvre donc potentiellement à toutes les langues et systèmes graphiques, d'abord ceux qui sont représentés dans les bibliothèques, archives, laboratoires et musées de France** (ce qui n'est pas limitatif), mais aussi plus largement à l'international, comme le fait déjà IIIF Collections.



Portail Biblissima : exemple d'interrogation sur le descripteur « dragons » (détail)

Qui dit toutes les langues et systèmes graphiques dit aussi tous les types de support. Pour certaines langues, le méroïtique par exemple, il n'y en a pas d'autres que des inscriptions antiques. L'alliance de IIIF et des fonctionnalités du visualiseur Mirador permettra la **comparaison de tous les artefacts porteurs de textes, en intégrant la 3D**. Cette évolution de Biblissima+, qui générera de nouveaux usages dans des communautés peu touchées aujourd'hui par l'équipex, celles des linguistes et des archéologues, épigraphistes et numismates en particulier, **va de pair avec la création, sur le Campus Condorcet, d'un nouvel Institut des langues rares (ILARA)**. Cet **institut créé par l'EPHE-PSL** grâce à une aide exceptionnelle du MESRI au titre du plan SHS 2020, en partenariat avec les laboratoires LLACAN et LACITO et l'INALCO, est consacré à la documentation et à la transmission des langues disparues documentées par l'écrit (une centaine de langues anciennes dont l'EPHE-PSL compte souvent l'un des rares spécialistes dans le monde, par ex. certaines langues très rares de l'Iran ancien, de l'Asie, de l'Afrique…) et des langues sans tradition écrite. C'est un centre de formation et de médiation, et donc un centre de ressources, tant documentaires que logicielles, travaillant étroitement avec la TGIR Huma-Num. Pour **l'ensemble des langues anciennes connues seulement par l'écrit**, il n'est pas utile que l'ILARA recrée une interface spécifique ou de nouveaux mécanismes d'acquisition des données : Biblissima+ sera la porte d'entrée vers les ressources

catalographiques, bibliographiques, et vers la numérisation des sources. L'alliance possible de Biblissima+, Persée, HAL et Collex en assurerait l'enrichissement et la qualité à l'échelle française, et, à travers une collaboration avec Istex et OPERAS, à l'échelle internationale (cf. infra).

Cette ouverture nouvelle de Biblissima+ lui permet d'associer pleinement des laboratoires ayant depuis longtemps intégré le multilinguisme des sources : non seulement **l'IRHT**, mais aussi le **CIHAM** (en particulier pour les sources arabes) et **HiSoMA**, qui travaille également sur des sources grecques, arabes, araméennes, latines et syriaques et dont un chercheur est l'un des PI du projet ERC Synergy **DHARMA** (ERC 809994) *The Domestication of "Hindu" Asceticism and the Religious Making of South and Southeast Asia* (2019–2025). Les outils déployés dans le cadre de **e-Scripta**, programme de recherche interdisciplinaire de PSL, sont destinés dès le départ à prendre en charge la diversité des langues, de leurs systèmes graphiques et de la mise en page de leurs textes.

… combler les souhaits de l'utilisateur …

A-I.2/ Atteindre l'exhaustivité sur les collections françaises

Implication : équipe Biblissima+, SIAF, IRHT, GED

Il n'est évidemment pas possible d'atteindre l'exhaustivité à l'échelle internationale, aussi l'ambition d'exhaustivité du projet ne concerne-t-elle pour les prochaines années que les fonds conservés en France, ce qui est déjà considérable, avec des accroissements au fil des partenariats pour les collections étrangères, comme Biblissima l'a déjà fait avec Wellcome Collection à Londres, et le fait régulièrement via IIIF Collections.

A-I.2.1/ Accès aux sources : partenariat avec les AN et FranceArchives

Implication : équipe Biblissima+, SIAF

Le travail de mise en interopérabilité des ressources des partenaires a parfois pris du retard. Par ailleurs, Biblissima a aidé des projets qui n'étaient pas prévus au départ, et dont les données ont dû ou doivent encore être agrégées aux autres ressources du portail. Chez les partenaires, certaines ressources ont été laissées volontairement en dehors du périmètre, et pourraient rejoindre l'infrastructure au cours des dix prochaines années si cela devenait pertinent scientifiquement. Mais le vrai problème de l'élargissement du spectre de Biblissima n'est pas là. Il concerne plus largement, pour Biblissima qui travaille beaucoup avec les services d'archives et les bibliothèques de France possédant des collections patrimoniales, **les fonds d'archives** (qui contiennent aussi souvent des manuscrits, mais aussi un nombre incalculable de documents intéressant l'histoire des bibliothèques) **et ceux des bibliothèques de France**.

L'internaute a besoin, s'il fait une recherche sur un texte, un auteur, un collectionneur, un lieu, une cote, une bibliothèque, un thème iconographique etc., que soient clusterisées par le portail toutes les informations concernant cet objet : toutes celles qui viennent des catalogues en ligne (en particulier ***Calames*** pour les collections patrimoniales des bibliothèques universitaires, et le *Catalogue général des manuscrits des bibliothèques publiques de France* (***CGM***) pour les autres bibliothèques [municipales, ministères, etc.], deux ensembles consultables à travers le *Catalogue collectif de France* [***CCFr***], sans compter les catalogues d'incunables rétroconvertis en grande partie grâce à Biblissima et désormais nativement numériques), toutes celles qui viennent des **bibliothèques numériques**, toutes celles qui viennent des **bases de données**, des **éditions électroniques**, que les sources soient des manuscrits, des imprimés, des documents d'archives, et quel que soit le type d'institution qui les conserve. Rappelons que la consultation de toute ressource clusterisée par Biblissima+ génère du flux pour le site source (soit directement soit parce que les internautes rebondissent vers la source des informations).

Biblissima+ va donc intégrer massivement des notices de n'importe quel réservoir de données en open access : par ex. celles du ***CCFr*** et toutes les données ouvertes de la **Bibliothèque nationale de France**, ce qui va changer la vie des chercheurs (Biblissima+ pourrait apporter ultérieurement des métadonnées et un niveau d'agrégation plus fin au CCFr). Côté archives, **FranceArchives** agrège déjà les données de 75 services du réseau des archives de France (les 3 SCN des Archives nationales, la Médiathèque de l'architecture et du patrimoine, le Service historique de la Défense, les Archives diplomatiques, 48 services d'archives départementales, 10 services d'archives municipales, 9 établissements publics et 2 associations) et a vocation à toutes les rassembler. Biblissima accompagne les **AN** dans leur passage à IIIF, en particulier à la suite du financement de la numérisation de séries d'inventaires médiévaux, modernes et d'époque révolutionnaire, et devrait accompagner le passage à IIIF des AD. Etant donné

l'ampleur des fonds conservés par les services d'archives, il serait raisonnable de cibler pour commencer certaines séries, par exemple les fonds des établissements religieux et des érudits, les documents permettant de reconstituer des fonds médiévaux. Le modèle des données de Biblissima concerne les anciens possesseurs (personnes physiques / morales) indépendamment du type d'entités qu'ils ont collectionnées, ce qui permettra de relier à des possesseurs aussi bien des archives, des emblèmes ou des sceaux que des manuscrits ou des imprimés. Ce sera un *work in progress*.

Nous avons en France tous les moyens de créer, à partir des ressources des bibliothèques, des archives, et sur ce modèle, un jour, des musées, **un portail du patrimoine des cultures écrites anciennes large et en même temps efficace**, qui pour le moment n'existe nulle part ailleurs. L'Italie, l'Allemagne, les Pays-Bas ont par exemple des portails nationaux, mais seulement pour les manuscrits médiévaux. En revanche, il faut accorder la plus grande attention à un nouveau projet britannique de grande envergure qui serait comparable et pourrait parfaitement décrire ce que nous pourrions faire pour les numérisations, *Towards a National Collection : Opening UK Heritage to the World*, qui s'appuie comme le nôtre sur le protocole d'interopérabilité des images IIIF :
« *Towards a National Collection is a major five-year £18.9 million investment in the UK's world-renowned* ***museums, archives, libraries and galleries****. The programme will take the first steps towards* ***creating a unified virtual 'national collection' by dissolving barriers between different collections*** *– opening UK heritage to the world. By seizing the opportunity presented by new digital technology, it will allow researchers to formulate radically new research questions, increase visitor numbers, dramatically expand and diversify virtual access to our heritage, and bring clear economic, social and health benefits to communities across the UK. The innovation driven by the programme will maintain the UK's world leadership in digital humanities and set global standards in the field. The Programme's main objectives are:*

- *to begin to dissolve barriers between different collections*
- *to open up collections to new cross-disciplinary and cross-collection lines of research*
- *to extend researcher and public access beyond the physical boundaries of their location*
- *to benefit a diverse range of audiences*
- *to be active and of benefit across the UK*
- *to provide clear evidence and exemplars that support enhanced funding going forward.* »

Certains domaines, comme la papyrologie (voir le remarquable portail Trismegistos) ou la diplomatique (voir par ex. Monasterium), ont des portails dédiés qui rassemblent leurs communautés. Mais **Biblissima+ va au-delà de ces initiatives par sa grande transversalité** culturelle, linguistique, chronologique, et parce qu'il intègre **non seulement des données patrimoniales, mais aussi toutes les données des unités de recherche, et donc leurs problématiques et leurs méthodes d'enrichissement et de réutilisation des données** : c'est pour cela que, du portail et des données qu'il rassemble, jailliront de nouveaux questionnements et de nouvelles découvertes.

A-I.2.2/ Pour un portail national, avoir un référentiel des cotes pour la France

Implication : équipe Biblissima+, IRHT, SIAF

Pour préparer cette évolution, Biblissima est en train de créer dans son référentiel de cotes de manuscrits, outil mis à disposition sur sa plateforme **data.biblissima** qui compte à ce jour près de 200 000 cotes, **l'ensemble des cotes des manuscrits des bibliothèques publiques et universitaires de France**, dont l'ID Biblissima va permettre l'agrégation des informations les concernant. **La même opération est en cours pour les incunables, et devrait être menée pour les cotes des archives**, qui représentent un énorme chantier à traiter progressivement (cf. supra). Les archives de chercheurs, souvent conservées dans des **laboratoires** et cotées seulement en partie, constituent un autre réservoir et un autre problème. C'est la constitution de ce référentiel, épine dorsale du projet Biblissima, qui a permis à Biblissima de suggérer pour la première fois l'idée d'un équivalent international de l'ISBN pour les manuscrits (mai 2017), reprise par le **consortium international ISMI**, porté en France par Biblissima et surtout l'IRHT, qui la développe très efficacement (et qui pourrait réutiliser le référentiel publié par Biblissima). Potentiellement, ce qui sera fait de façon exhaustive pour la France pourrait l'être pour des fonds étrangers, au moins pour les manuscrits, en particulier à travers IIIF Collections et ISMI (pour les incunables, une association avec la base de données MEI Material Evidence in Incunabula, créée par Cristina Dondi à Oxford, ferait sens).

L'association progressive de jeux de données en langues anciennes transmises par autre chose que des manuscrits de type occidental, des documents d'archive ou des imprimés, comme les papyrus et *ostraka*, les supports de toute sorte (tablettes d'argile, feuilles de palmier, écorces de bouleau…), les objets issus de fouilles archéologiques, le matériel épigraphique, les sceaux, les monnaies etc. imposeront, au fil de la vie de Biblissima+, **l'intégration d'autres systèmes d'identification, comme les numéros d'inventaires de musées** par ex. Sur les référentiels en général, en particulier sur l'alimentation du FNE, voir infra, A-II.1.

A-I.3/ Donner un accès ciblé aux résultats de la recherche en articulant les programmes nationaux les uns avec les autres

Implication : équipe Biblissima+, GED, IRHT

Le champ de la bibliographie liée aux objets du portail, dont chaque catégorie fait l'objet d'un référentiel publié progressivement sur la plate-forme data.biblissima, n'existe pas, à ce jour, dans le projet Biblissima, si ce n'est indirectement, à travers les bases de données, les catalogues, les éditions en ligne et la bibliographie à laquelle ils renvoient. On accède ainsi à une partie de l'information, mais les ressources bibliographiques proprement dites (articles, ouvrages) ne sont pas présentes activement. Or on voit facilement quel enrichissement cela représenterait pour l'internaute de chercher par exemple une cote de manuscrit, ou un nom de personne, et de recevoir des suggestions bibliographiques **avec des liens vers les ressources en ligne et vers les bibliothèques qui en sont détentrices, en limitant le « bruit » grâce à l'usage des ID**. C'est en fait un enjeu colossal du point de vue de la recherche, car on peut essayer de faire une telle enquête sur internet avec un moteur de recherche, mais c'est très compliqué à cause des formes multiples de tous les objets historiques concernés (y compris la multiplicité des formes liée au multilinguisme), ce qui génère un « bruit » considérable, beaucoup de perte de temps, et n'offre aucune garantie de succès ou de qualité. C'est l'alignement autour d'un ID unique qui va permettre de réunir des informations pertinentes autour d'un objet et d'effectuer une recherche de qualité.

Un précédent européen peut nous aider à illustrer cet enjeu : en Allemagne, la Deutsche Nationalbibliothek (DNB) a convaincu des projets allemands, suisses et autrichiens d'adopter son référentiel, ce qui permet d'aligner toutes les ressources sur les manuscrits de ces trois pays, en particulier pour les noms de personnes. Cela permet à la plate-forme *Regesta Imperii* de réunir une bibliographie indispensable pour toutes les études médiévales – que nous pouvons d'ailleurs récupérer via l'alignement de nos propres ID Biblissima avec ceux de la DNB. Si les bases de données de recherche, les catalogues, les bibliothèques numériques etc. de France disposaient d'un tel alignement, en particulier grâce au FNE, nous pourrions en faire autant et plus.

La France s'est dotée d'un ensemble d'outils remarquables : Persée et Istex pour les archives de la littérature scientifique, Collex pour la numérisation des collections et archives de la recherche, HAL pour l'archivage ouvert des résultats de la recherche, OpenEdition qui, par son alliance avec la TGIR Huma-Num, a créé le consortium OPERAS à l'échelle européenne pour les ressources bibliographiques vivantes. **Biblissima+, constatant l'usage important de ces outils par les communautés qu'il fédère, propose de démontrer comment ces différents outils de grande envergure s'articulent et s'enrichissent mutuellement** sur la thématique large et variée de l'histoire des cultures écrites anciennes.

Biblissima+ peut devenir pour tous ces programmes le **lieu de traitement de toutes les données complexes, souvent floues ou incomplètes** concernant les entités en jeu dans l'histoire de la transmission et de l'étude des cultures écrites, et **leur fournir ses référentiels spécialisés avec leurs alignements vérifiés**. Ceux-ci leur permettront de créer des pages d'agrégation des données autour des entités vérifiées par Biblissima+, et Biblissima+, de son côté, donnera accès à ces ressources bibliographiques. Les liens réciproques permettront d'agréger les données de ces grands acteurs du monde des bibliothèques, des archives et de l'édition avec celles du monde de la recherche et de l'enseignement supérieur.

A-I.3.1/ Littérature scientifique : préparer un partenariat avec Persée, HAL, et Istex

Implication : équipe Biblissima+, GED

La France s'est dotée d'une infrastructure de numérisation et de mise à disposition de ses revues et collections SHS, **Persée, dont les ressources sont particulièrement pertinentes à propos du patrimoine écrit des bibliothèques et archives de France, et plus généralement de l'histoire des cultures écrites**, et accessibles à n'importe quel internaute français ou étranger. Les objectifs de Biblissima+ sont partagés par Persée, qui développe également des outils de recherche sémantique (http://data.persee.fr). Une alliance de Persée et de Biblissima+, avec un partage des référentiels, en particulier ceux des cotes, des noms de lieux historiques, des noms de personnes et d'institutions (souvent très rares dans le cas de Biblissima+), **lèverait un verrou énorme et constituerait pour les chercheurs une avancée impressionnante**. Dans le portail Biblissima+, l'interrogation sur les objets renverrait aux ressources de Persée et permettrait de constituer sur chaque objet une documentation très riche. Dans Persée, la recherche bibliographique permettrait de rebondir grâce au portail Biblissima+ vers les sources primaires et les dernières avancées de la recherche les concernant afin de mettre en perspective la bibliographie et de la dépasser. On peut imaginer, dans un second temps, des recherches d'une autre nature, **par exemple reconstituer la documentation de tel savant, ou suivre les développements de l'histoire des idées et de la critique à propos de tel concept, de tel personnage, de tel texte** et ainsi de suite, en particulier à partir des comptes rendus. Cette ressource est très complémentaire de Gallica, où l'on

trouve aussi quantité de collections et revues numérisées. Persée ne peut pas s'engager dans ce travail dès 2021, car sa priorité est de disposer de sa nouvelle plateforme, qui, selon l'obtention ou non de financements, sera disponible dans 2 ou 4 ans. **Le balisage proposé par Biblissima+ ne pourra être mis en œuvre qu'après, dans la deuxième moitié de Biblissima+,** qui aura enrichi ses référentiels dans l'intervalle. Dans les ressources de Persée, ont été identifiées à ce jour **41 revues et collections** sur lesquelles porterait l'effort, dont 26 volumes du *Corpus des inscriptions de la France médiévale* (https://www.persee.fr/collection/cifm), qui couvrent les trois quarts de la France (pour les publications les plus récentes, voir le projet d'édition en TEI sur TITULUS), et les principales revues des partenaires de Biblissima+. Des lettres d'intention ont été échangées entre Biblissima+ et le projet POPS porté par l'ENS de Lyon pour Persée.

Il serait également très utile de fournir l'ensemble des référentiels de Biblissima+ à HAL pour parvenir à un **balisage fin et exact des entités nommées**, au moins dans les résumés, tables des matières, index, afin d'aller au-delà de l'analyse des PDF et leur restructuration automatisée grâce à l'outil Grobid. Cela permettrait à Biblissima+ et donc aux chercheurs de **moissonner HAL de façon fine** afin de valoriser les métadonnées de cette archive ouverte nationale, et, par contre-coup, de s'associer au vaste mouvement d'encouragement à son alimentation par les communautés de chercheurs humanistes. Cependant, la priorité de HAL est aujourd'hui de renouveler sa plateforme en profondeur (projet d'équipex+ HALliance), et le projet proposé par Biblissima+ (améliorer les métadonnées liées aux articles de HAL par un alignement sur les référentiels data.biblissima et, du côté de HAL, établir des liens avec des contenus numérisés) vient trop tôt. Le CCSD et Biblissima+ s'entendent pour travailler plutôt à une réponse commune à d'autres appels à projets.

Enfin, de la même façon, il serait également très intéressant pour les internautes de pouvoir **accéder directement**, à partir toujours des référentiels, **aux suggestions d'Istex**, énorme réservoir d'archives électroniques de publications achetées par la France grâce au PIA, et qui s'est doté de puissants outils de TDM (*Text and Data Mining*) permettant d'explorer les ressources bibliographiques avec une granularité fine (accès aux sujets, aux articles, aux entités nommées etc.). Il n'y a à cela qu'une limite : à ce jour, les ressources Istex ne sont accessibles que via la fédération d'identités Renater, l'accès au texte intégral est donc réservé au public de l'ESR, pour lequel les droits ont été négociés. Au moins pour les usagers de l'ESR, on rentabiliserait ainsi de façon très élégante des produits financés séparément grâce aux Investissements d'avenir, en les articulant plus finement encore les uns aux autres sur le volet qui nous concerne. Pour les autres, Istex fournirait au moins des références bibliographiques. **Persée, HAL et Istex donneraient accès à une superbe bibliothèque virtuelle**, dont les ressources seraient ainsi directement accessibles depuis notre portail.

A-I.3.2/ Négociations avec des éditeurs : BAMAT-O et e-Scriptorium via l'IRHT

Implication : IRHT, équipe Biblissima+

L'IRHT peut apporter sa contribution à cette partie bibliographique de Biblissima+ par le biais de la **BAMAT-O** et d'**e-Scriptorium**.

La BAMAT-O est la version électronique de la *Bibliographie annuelle du Moyen Âge tardif*, préparée à l'IRHT depuis 1991 et publiée par Brepols. L'ouverture de la BAMAT-O en 2020, fruit d'un partenariat renouvelé entre l'IRHT et Brepols, propose en ligne l'ensemble de cette bibliographie critique par les entrées Auteurs anciens, Auteurs modernes, Œuvres, Thèmes, Manuscrits. La base est payante sur le bouquet Brepolis, mais un accès direct se fera depuis la base Medium à partir des cotes des manuscrits, jusqu'à un certain niveau d'information ; une négociation en cours avec Brepols devrait permettre un accès plus profond dans les années à venir.

La même démarche est adoptée avec eScriptorium, qui rassemble les notices du *Bulletin codicologique* (600 à 700 notices par an, multilingues) de la revue *Scriptorium. Revue internationale des études relatives aux manuscrits* (fondée en 1946) et indexe les articles de la revue par les cotes des manuscrits. Depuis quelques années, grâce au soutien du consortium COSME puis COSME^2, l'IRHT et la revue travaillent à l'alignement et à la mise en relation avec Medium des quelque 250 000 cotes de manuscrits concernées par cette publication. L'accès libre via Medium ouvre ainsi le champ de l'essentiel de la littérature critique sur les manuscrits latins, français et grecs jusqu'à 1500.

A-I.3.3/ Accéder aux revues et ouvrages vivants : vers un partenariat avec OPERAS

Implication : équipe Biblissima+, GED

Dans la même logique, **Biblissima+ prévoit de collaborer avec la TGIR Huma-Num et Open Edition / OPERAS pour affiner les outils de découverte** en ce qui concerne toutes les données produites dans les domaines des sciences de l'érudition, et en retour, pour fournir à ses usagers un accès direct aux publications ouvertes non seulement d'OpenEdition, mais aussi de ses partenaires internationaux. En sens inverse, comme avec Persée, OPERAS pourrait utiliser les alignements non seulement pour la fouille de ses données, mais aussi pour donner à ses lecteurs un

accès direct aux données de la recherche clusterisées par Biblissima+. OpenEdition, qui est en charge de l'accompagnement des fournisseurs de données dans la plateforme, a d'ores et déjà **ajouté Biblissima/Biblissima+ à la liste des fournisseurs potentiels de données pour le projet TRIPLE**, et l'équipe d'Huma-Num travaillera avec Biblissima+ à une possible mobilisation de ses référentiels dans la plateforme. Biblissima+ s'engage à intégrer le « comité de partenaires » du projet d'équipex+ COMMONS présenté par Huma-Num, OpenEdition et Metopes, qui a pour objectif de traiter l'articulation entre la gestion des données de recherche et les publications dans le cadre de la source ouverte. Ce projet intéresse fortement les communautés scientifiques mobilisées au sein de Biblissima+. Cette proposition s'inscrit dans la logique de cycle de vie des données et donc de bonne articulation des acteurs les uns avec les autres, de la production des données jusqu'à leur publication et leur réutilisation lors d'un nouveau cycle (cf. *infra*, p. 25). Des lettres d'intention ont été échangées entre les deux projets Biblissima+ et COMMONS.

A-I.3.4/ Liens avec les bibliothèques physiques : le GED et le SUDOC

Implication : équipe Biblissima+, GED

Il reste important de **localiser les ouvrages imprimés**, ne serait-ce que pour se mettre en relation avec les bibliothèques détentrices et encourager leur numérisation. Ce qui serait utile ici, ce serait un interfaçage via les référentiels avec le SUDOC et la BnF.
Un travail spécifique avec le **GED** permettra de travailler à un alignement des identifiants et des référentiels entre Biblissima+ et la base de données du GED. Une des exploitations possibles pourrait être par exemple la création de liens depuis les références bibliographiques attachées aux notices de manuscrits exposés dans le portail Biblissima+, vers les numérisations et les notices des ouvrages cités dans les bases de données. Ce travail mené avec le GED servira de preuve de concept pour élargir cet alignement aux bibliothèques de l'ESR qui souhaiteraient s'inscrire dans la même démarche.

Par ailleurs les capacités des outils considérés, de part et d'autre, à exposer leurs métadonnées sur le web de façon fine et sémantique permettront de valoriser de façon croisée les informations et ainsi de rapprocher bibliothèque physique et bibliothèque numérique.

A-I.4/ Naviguer dans les archives de la recherche et les bibliothèques numériques patrimoniales de l'ESR

Implication : équipe Biblissima+, GED, en lien avec Collex-Persée

L'expérience de Biblissima, et tout particulièrement des historiens des bibliothèques anciennes, démontre que pour écrire l'histoire culturelle de l'Antiquité, du Moyen Âge, de la Renaissance, il faut très souvent avoir recours aux archives de l'époque moderne, jusqu'à un XIX^{e} siècle avancé, et même davantage, puisqu'il est souvent crucial de pouvoir suivre les pérégrinations d'un document jusqu'à la collection contemporaine qui le conserve, publique ou privée. Biblissima+ souhaite donc donner à cette documentation moderne et contemporaine plus de visibilité et d'espace. Le référentiel des cotes inclura donc **l'ensemble des cotes de manuscrits et archives des établissements de conservation, sans distinction de date**. Comme dans Biblissima, les campagnes de numérisation pourront inclure de tels documents, ou même les concerner exclusivement.

Dans ce contexte, **une collaboration avec Collex-Persée et les bibliothèques numériques patrimoniales des bibliothèques universitaires** prendrait tout son sens, avec une mise en relation systématique des archives de la recherche et de leurs sources manuscrites ou imprimées. Le GED pourrait porter plus particulièrement ce volet de Biblissima+ et assurer le lien avec les bibliothèques délégataires de Collex-Persée.

Biblissima partage déjà les mêmes grands principes que Collex, et fait travailler ensemble, depuis 2012, chercheurs, formateurs, conservateurs, ingénieurs pour parvenir à développer une politique de numérisation des documents médiévaux et modernes fondée sur la recherche, signifiante, à même de créer des communautés de recherche. Sa position d'auxiliaire, d'observateur et de fédérateur lui permet aujourd'hui de faire un constat : autant on voit les campagnes de numérisation des bibliothèques municipales converger vers trois grandes bibliothèques numériques, la BVMM (IRHT, CNRS), les BVH (CESR, Tours) et Gallica (BnF), avec une tendance actuellement à faire de Gallica, la moins spécialisée d'entre elles, la grande bibliothèque fédératrice de ces bibliothèques municipales, autant **les choix de l'enseignement supérieur et de la recherche ont privilégié la mise en place d'outils spécifiques, adaptés aux situations locales,** ce qui rend impossible une recherche transversale dans les corpus numérisés, qui pourtant permettrait des rapprochements propres à susciter des découvertes.

Dans l'objectif d'une meilleure fédération des usages, Biblissima+ continue de favoriser l'adoption du protocole d'interopérabilité des images IIIF dans les

bibliothèques et projets de l'ESR, comme il le fait depuis deux ans avec le service IIIF 360, en partenariat avec la TGIR Huma-Num et le Campus Condorcet. Dans le contexte de Biblissima+, le portail Biblissima+, en partenariat étroit avec le GED, proposera **un « paquetage » de modules Omeka que IIIF 360 pourra contribuer à maintenir, à développer, à documenter**, tout en assurant **l'accompagnement**, **l'analyse des besoins** en amont etc. Ces modules concerneront à la fois la partie « serveur IIIF » (images et manifests), la partie « visualisation IIIF » (Mirador), éventuellement un module d'annotation, et aussi les aspects métadonnées (alignement avec les référentiels, moissonnage et exposition des métadonnées etc.). Il y aura dans tout cela un lien naturel avec l'outil **NumaHop** mis au point par des BU pour la gestion des campagnes de numérisation, l'acquisition et le partage des images avec les prestataires, le contrôle qualité etc., dont est également en train de se doter le GED (et qui fait partie du projet POPS), et, à la TGIR Huma-Num, **Nakala** pour le dépôt et la publication des images, l'exposition via IIIF.

Avec la mise à disposition des référentiels et dans la logique de IIIF Collections, Biblissima+ donnerait les moyens de créer et d'enrichir **un lieu d'agrégation des données des bibliothèques numériques patrimoniales de l'ESR**. Pour l'outil Primo adopté par plusieurs bibliothèques dernièrement, il serait utile de multiplier les alignements entre SUDOC et Calames (qui existent déjà mais ne sont pas assez larges) pour maximiser la *discoverability* des fonds patrimoniaux des bibliothèques numériques de l'ESR. Celles-ci y gagneraient du flux d'internautes, pour quatre raisons : l'une technique, le protocole IIIF permettant une consultation dynamique dans l'interface de consultation, mais toujours décomptée chez le détenteur de la ressource ; les trois autres liées au comportement général des internautes, qui se tournent plus volontiers **vers une grande bibliothèque numérique simple d'usage**, qui ne va pas, pour chaque fonds, leur demander un réapprentissage, **qui va favoriser la sérendipité** de la recherche, de plus en plus appréciée et valorisée aujourd'hui, et qui va permettre, **grâce aux fonctionnalités développées par le portail Biblissima+ et à l'ampleur de ses ressources, des rapprochements inédits et la visualisation simultanée des documents**.

Le GED pourrait prendre sous sa responsabilité l'inscription dans le cadre des appels à projets Collex-Persée de cette problématique de fédération des contenus modernes et contemporains.

A-I.5/ Etablir le lien avec l'édition des textes

Implication : équipe Biblissima+, IRHT, eScripta

Comment faire un pas de plus dans l'aide aux chercheurs, mais aussi dans la mise à disposition de tous des traces écrites des cultures anciennes ? L'analyse des pratiques de recherche actuelles fait apparaître le besoin que soit lié au manuscrit, à l'inscription, à l'objet portant une écriture quelle qu'elle soit, **une transcription ou une édition critique**, et parfois, si c'est possible, une **traduction**.

Grâce aux outils d'OCR avancé décrits dans la partie B, on pourrait imaginer, **pour des internautes novices**, qui abordent par exemple les écritures manuscrites mais ne savent pas encore bien les déchiffrer, un **affichage à la volée de la transcription** du document qu'ils regardent en ligne. On dépasserait ainsi, pour le grand public, le niveau de la frustration, mais on rendrait aussi un grand service à tous ceux qui, par exemple, font de la **généalogie** sans avoir été formés à la lecture des écritures anciennes. La transcription à la volée, par exemple au passage de souris, pourrait s'accompagner, pour ce public, d'une traduction ou même d'une **lecture à haute voix** du texte édité (quand il l'est) synchronisée avec la division en pages du manuscrit. Pour ce faire, il pourrait être intéressant de conclure un accord avec un éditeur ou des éditeurs pour toute la littérature classique et tardoantique et quelques textes médiévaux. Il serait intéressant aussi de proposer un volet de traductions bilingues : des propositions en ce sens pourraient apparaître en réponse aux aap annuels.

Pour les chercheurs, l'ensemble des équipes impliquées dans Biblissima développent des éditions électroniques de sources, dans divers environnements qui seront présentés en partie B, et les ont liées (ou souhaitent le faire) avec leurs bases de données documentaires : sources historiques, diplomatiques, littéraires, exégétiques, épigraphiques (de l'Antiquité grecque à l'époque moderne, en passant par le *Corpus des inscriptions de la France médiévale*, les sceaux, et l'épigraphie de la Renaissance). Le projet d'infrastructure numérique Biblissima+ ne prévoit pas de financer tel corpus plutôt que tel autre, mais prévoit des **mécanismes d'acquisition**. Le premier mécanisme est celui qui va permettre d'enrichir l'équipement de données de la façon la plus pertinente en corpus balisés en XML-TEI, à travers le financement de projets partenariaux annuels (ci-dessous). Le second mécanisme consiste à développer des outils de transcription automatisée utilisant l'intelligence artificielle, décrits dans la partie B. L'infrastructure, qui va développer les possibilités d'annotation des sources numérisées, va également faciliter la constitution **d'environnements de transcription et de comparaison** des artefacts anciens et des textes édités (les liens vers les très nombreuses éditions en ligne seront systématisés), ainsi que de leurs **diverses versions**.

A-I.6/ Fédérer la diversité des initiatives : développer les appels à projets partenariaux

Implication : équipe Biblissima+, GED

Quantités d'idées de projets de courte ou moyenne durée, sur des corpus ou des outils précis, ont été exprimées au fil des années et pendant la préparation du projet. A elles toutes, elles ne font pas une infrastructure, mais combinées avec l'ensemble des mesures proposées dans Biblissima+ et structurées par ses outils et l'exigence d'adoption de formats ouverts et d'interopérabilité, elles feraient de cette infrastructure **un bel ensemble interconnecté, solide dans sa structure générale et riche dans le détail**, au service de toutes les communautés. La meilleure réponse à ces besoins, modulable dans le temps, donc garantissant de pouvoir **répercuter ou créer les innovations techniques et scientifiques pendant 8 ans**, nous semble le système des appels à projets annuels.

Les appels à projets partenariaux de Biblissima ont été l'une de ses plus grandes réussites. Le principe était de **favoriser la collaboration entre institutions de conservation et équipes de recherche** en conditionnant l'obtention des fonds pour de la numérisation, du catalogage, le développement d'outils à l'existence d'un projet scientifique, commun à une équipe de recherche et à l'établissement de conservation porteur du projet. Les financements étaient confiés à des partenaires de Biblissima, qui garantissaient la bonne exécution du projet. Les projets partenariaux ont apporté un souffle nouveau aux bibliothèques et archives, et favorisé une politique scientifique de numérisation des documents anciens. Scientifiquement, ils ont été la part d'imprévu, et donc d'inventivité du projet. Ils ont par ailleurs créé un **maillage du territoire en l'irriguant à l'aide des fonds équipex**, de telle sorte qu'une soixantaine de bibliothèques, archives, équipes de recherche ont bénéficié de l'aide de Biblissima. C'est aussi par ces projets partenariaux que se sont mises en place des opérations de formation, **les écoles d'été Biblissima, grâce à l'effet-levier des financements** : tout projet ayant reçu au moins 40 000 € devait organiser par ses propres moyens une école internationale ; 120 étudiants et professionnels français et étrangers y ont été formés. Nombreux sont ceux qui, aujourd'hui, travaillent dans les projets d'humanités numériques, dans les fonds patrimoniaux, dans l'enseignement supérieur et sont passés par ces écoles, qui ont joué un rôle déterminant en leur démontrant **la fécondité et donc la nécessité de l'alliance des disciplines de l'érudition et de la pratique du numérique**.

Les communautés Biblissima

Biblissima+, contrairement à Biblissima, n'a plus besoin de financer des projets « fondateurs », car le socle scientifique est désormais acquis pour l'essentiel. En revanche, **il reste vital de prévoir un financement pour l'innovation et la formation**, qu'il s'agisse de projets scientifiques inventifs producteurs de données nouvelles, de nouvelles ressources à mettre en interopérabilité via le portail, ou de nouveaux outils et services pour la recherche, dans un esprit d'ailleurs commun à Biblissima+ et à Collex. L'environnement de travail de Biblissima+ a vocation à mettre à disposition les outils ainsi produits (sans exclusivité) pour en favoriser l'adoption par les communautés. C'est **d'autant plus vital que le portail s'ouvre à de nouveaux domaines linguistiques**. Par ailleurs, **l'extension du périmètre de Biblissima+ aux services d'archives**, et en particulier aux services d'Archives départementales, offre l'opportunité de mettre en valeur leurs riches fonds médiévaux, les inventaires révolutionnaires ainsi que les inventaires après décès et d'en favoriser la description et la numérisation dans des formats directement exploitables par le portail. Les enjeux resteront les mêmes : favoriser la production de données ouvertes et nativement interopérables, leur mise à disposition à travers des ressources numériques existant déjà, accessibles par le portail Biblissima+, et **former de nouvelles générations de chercheurs et de professionnels des bibliothèques, archives et musées**, mais l'ouverture linguistique, temporelle et documentaire donnera à ces projets un nouvel intérêt. Tout projet de numérisation, de production de données, de développement d'outil comportera l'obligation de **contribuer au portail, le temps de travail et les développements nécessaires pour cela** étant inscrits dans le projet. Ces objectifs justifieraient au moins un doublement de la somme annuelle (200 000 € / an de 2013 à 2018).

Il faudrait aussi, pour répondre à une demande que Biblissima a souvent reçue, **ouvrir ces projets partenariaux, dans une certaine proportion, à des collaborations internationales**. Il est rare, étant donné l'histoire du collectionnisme et des fonds modernes, de pouvoir constituer un corpus scientifique (par exemple reconstituer une collection médiévale) sans faire appel à des documents dispersés à travers le monde. D'autre part, le dispositif suscite l'intérêt à l'international, comme le montrent les réactions enthousiastes lors des présentations de Biblissima dans des colloques, et crée beaucoup de désirs de collaboration. La plateforme IIIF Collections permet de répondre en partie au désir des bibliothèques étrangères de rendre leurs collections cherchables via le portail, mais leur offrir la possibilité de financer des campagnes de numérisation ou de catalogage pertinentes pour l'Observatoire des cultures écrites anciennes serait un grand pas en avant. Nous proposons de limiter à 20% du total cette part dévolue aux opérations internationales.

Dans le nouveau paysage documentaire structuré depuis la création de l'infrastructure de recherche Collex-Persée, **les appels à projets financés par Biblissima seront mis en lien avec les appels à projets portés par le GIS Collex,** afin de maximiser le retour sur investissement des sommes engagées.

… une colonne vertébrale : les référentiels …

A-II – Investir dans la durée : référentiels et mises à jour

A-II.1/ Les référentiels Biblissima+ à l'échelle nationale et internationale

Implication : équipe Biblissima+, IRHT, et l'ensemble des partenaires

Biblissima+ étant un projet d'infrastructure d'agrégation et d'enrichissement des données, l'accroissement constant en quantité et en qualité de ses référentiels est un élément clé de la réussite de l'infrastructure. Ils ont été évoqués à plusieurs reprises, puisqu'il y a **un lien étroit entre la qualité du travail d'agrégation et de mise en interopérabilité et celle des référentiels** : il ne suffit pas d'affecter à toute entité un identifiant, il faut aligner les données, les dédoublonner ou au contraire les fondre ensemble, les désambiguïser, travail particulièrement délicat dans toutes les disciplines de haute érudition, en particulier sur les textes en langues rares. Data.biblissima, avec tous ses partenaires, a ainsi un rôle à jouer non seulement pour le projet Biblissima+, mais aussi pour tous les grands outils utilisant des référentiels.

A-II.1.1/ Référentiel de cotes de manuscrits, d'archives et d'imprimés anciens (voire plus)

Implication : équipe Biblissima+, GED, IRHT, SIAF et AN, consortium ISMI

Biblissima a déjà créé dans son référentiel data.biblissima environ 200 000 cotes de manuscrits et d'incunables, et poursuit son travail

systématiquement pour la France : Biblissima+ devrait lui permettre de passer à une échelle supérieure en intégrant au moins une partie des cotes d'archives et en tendant à l'exhaustivité pour les collections françaises. La difficulté n'est pas sous-estimée : extrême diversité des systèmes, séries non-cotées (dans des archives, dans des laboratoires…), mais c'est un mouvement qu'il faut lancer.

Pour les établissements étrangers, l'intégration des cotes dans le référentiel se fait au fil de l'arrivée des données dans les ressources des partenaires, et, beaucoup plus massivement, au fil de l'accroissement des ressources de la plateforme IIIF Collections. Petit à petit se construit **un grand référentiel des livres anciens dans le monde**, qui devrait être élargi au moins aux archives, aux papyrus et *ostraka*, et probablement un jour aux autres objets porteurs de texte identifiables par leurs numéros d'inventaire. Pour les manuscrits, le **consortium international ISMI,** coordonné en France par l'IRHT et Biblissima, en Allemagne par la Bayerische Staatsbibliothek, en Suisse par E-Codices et l'Université de Fribourg, a une ambition plus élevée grâce à la fédération des établissements de conservation à l'échelle mondiale. Il serait possible de financer l'interface de consultation et d'interrogation d'ISMI, en prenant data.biblissima pour socle de départ et comme plateforme de stockage / gestion du référentiel, afin de ne pas créer de nouveau doublon. Pour les imprimés anciens, une collaboration avec la base de données MEI et le CERL serait particulièrement bienvenue.

A-II.1.2/ Développement des référentiels d'objets et concepts rares pour le FNE

Implication : équipe Biblissima+, GED, tous les partenaires

Pour les **noms de lieux et les noms de personnes (physiques et morales)**, Biblissima est spécialisé dans les noms rares liés au monde des textes et livres anciens du 5e au 18e s., pour lesquels on a parfois très peu de données, et qui sont en général pas ou mal gérés par les grands référentiels nationaux. Biblissima+ va permettre un accroissement considérable de ces référentiels, car cette caractéristique se retrouve dans le référentiel des noms de personnes des **AN** et des **archives** en général, tous les travaux sur les cartulaires, riches en entités nommées, menés à **l'IRHT**, en particulier le projet HOME (History of Medieval Europe) en lien avec la base CartulR, ou à **l'ENC**, dans les données prosopographiques et géographiques des différents programmes du **CESR** (catalogues d'incunables, projets d'éditions de textes, base de typographie de la Renaissance BaTyR, base Renumar : ressources numériques pour l'édition des archives de la Renaissance), dans la base prosopographique liée au laboratoire d'édition et d'annotation de sources de **Caen**, dans les bases de données de **l'EHESS** consacrées aux *exempla* et aux *quodlibeta*, dans le projet *Dictionnaire topographique de la France* (https://dicotopo.cths.fr) piloté par le **CTHS** à **l'ENC**, mené avec les **AN** et le **SIAF**, dans les référentiels collectifs de noms de lieux et de noms de personnes en cours d'élaboration par **FranceArchives**, dans le projet **e-Signa de l'EPHE-PSL** consacré aux marques héraldiques et aux sceaux, dans les bases de données de **l'IRHT** et les travaux du consortium d'Huma-Num **Cosme²** sur les référentiels d'institutions ecclésiastiques anciennes, dans les données de Biblindex, index en ligne des citations bibliques dans la littérature patristique de l'UMR **HiSoMA**, qui pourrait s'articuler avec les programmes Gloss-e (Glose biblique) et SourcEncyMe (corpus et sources des encyclopédies médiévales) de **l'IRHT** pour le traitement des citations d'auteurs et d'oeuvres. Avec l'apport de documents dans d'autres langues, dont beaucoup sont des langues rares concernant des documents difficiles, cette fonction va s'amplifier.

Biblissima a choisi pour son référentiel le modèle WikiBase parce que c'est celui que la France a choisi pour le Fichier National d'Entités (FNE) piloté par l'ABES et la BnF. Il est donc logique que la plateforme **data.biblissima, qui fédère des données produites de façon non concertée par les partenaires scientifiques**, soit fournisseur du FNE, et elle s'y est préparée. Biblissima+ propose de donner à data.biblissima un rôle de **« bac à sable » ou de « purgatoire » pour les données** de ce type produites par les projets de recherche ou des recherches individuelles (par ex. des recherches doctorales), afin **d'alimenter efficacement le FNE**, mais aussi les **référentiels internationaux** comme ISNI ou Geonames, en données de qualité, alignées de façon stable sur les grands référentiels (du type BnF, Library of Congress ou DNB pour les noms de personnes, les collectivités, les œuvres [plutôt que VIAF, qui est instable]). **Cette fonction s'inscrit dans le projet CPER Condornum du Campus Condorcet** (2021-2027), qui propose que Biblissima, le GED et Huma-Num s'associent pour créer en amont du FNE un « ***service standard de réconciliation de données***, *pour permettre l'alignement manuel ou semi-automatique de données existantes vers ces mêmes référentiels, en s'appuyant sur des outils d'IA en particulier* ». Ce travail pourra être réalisé en liaison avec les **Archives de France** puisque le portail FranceArchives a lui aussi vocation à servir de connecteur avec le FNE.

Pour les **descripteurs iconographiques**, Biblissima a déjà réalisé l'alignement des vocabulaires différents utilisés par les deux grandes bases de données françaises du domaine, Mandragore (BnF pour ses propres manuscrits) et Initiale (IRHT pour les autres manuscrits des bibliothèques de France), Initiale venant d'être intégrée au portail (juillet 2020). Cela va changer la vie de tous les utilisateurs de ces bases, qui devaient jongler avec des vocabulaires et donc des

approches différentes, pour reconstituer des corpus unifiés. C'est une première à l'échelle internationale aussi, car **Europeana n'a pas investi dans ce domaine**. D'autres thesauri iconographiques pourront être pris en compte, comme celui publié par l'EHESS (le TIMEL : https://datu.ehess.fr/timel). Dans le cadre de Biblissima+, les descripteurs seront **alignés de façon systématique avec Iconclass et wikidata** afin de permettre une interopérabilité plus large au niveau international et de surmonter ainsi la barrière des langues qui rend pour l'instant difficile l'interopérabilité en matière de vocabulaires iconographiques.

Les référentiels ne se limitent évidemment pas à ces grands domaines. L'IRHT a développé un vocabulaire codicologique multilingue, Codicologia, et songe à développer, à partir des descripteurs de la base de données Bibale, un vocabulaire de la **reliure** coordonné avec celui qui a été développé en anglais par *Ligatus* (N. Pickwoad, Londres). L'IRHT aimerait réaliser un alignement pérenne et automatisé de référentiels entre bases de données dans le **domaine des manuscrits grecs et orientaux**, à la fois celles qui sont développées par l'IRHT (*Pinakes* et *RIMG* pour le grec, *e-ktobe* pour le syriaque, *Medium* pour les manuscrits latins) et celles d'autres partenaires de Biblissima+ (en particulier HiSoMA) dans les mêmes domaines, travail déjà largement préparé par l'équipe Biblissima pour l'intégration de *Pinakes* dans le portail. Dans le cadre du consortium d'Huma-Num Cosme[2] a été menée une réflexion sur les référentiels de **valeurs et mesures**, de modes de **datation**, de **formulaires diplomatiques**. **Dates et concepts** feront partie du balisage mené dans le cadre de l'environnement de « lecture enrichie » conçu à l'EHESS, « Savoirs ». Les équipes impliquées dans la lemmatisation développent des référentiels **linguistiques**. Enfin, l'ENC et l'équipe Biblissima vont mettre au point ensemble, dans le cadre d'un projet de valorisation des thèses de l'ENC, un référentiel de **concepts historiques et historiographiques** qui pourrait être très utile pour la recherche dans Persée, HAL, Istex, OPERAS.

A-II.1.3/ Vers un multilinguisme des référentiels

Implication : équipe Biblissima+, IRHT

Biblissima+, comme le fait déjà Biblissima à moindre échelle, va intégrer de plus en plus de données issues de projets développés hors de France, car son modèle respectueux de la structuration native des ressources est attractif et la demande est croissante ; cela répond à une logique scientifique autant qu'à la logique du web. Cette logique de croissance nécessite de penser un **développement international des référentiels**, intégrant le multilinguisme. Le référentiel des cotes est le plus facile à mettre en place à l'échelon international justement parce que le problème de la traduction se pose moins. **L'initiative internationale ISMI** va dans ce sens. Pour d'autres vocabulaires, comme ceux de l'iconographie, de la paléographie, de la codicologie, le multilinguisme est en revanche un problème extrêmement sensible. L'expérience Biblissima montre que les données complexes et floues exigent **un travail de diplomatie scientifique et d'approfondissement qui va très au-delà de l'informatique**. Il nous semble que seule une structure internationale serait légitime pour porter des référentiels vraiment internationaux et la logistique nécessaire, et c'est pourquoi cet objectif sera inscrit dans un projet européen, Scriptissima, préparé en 2020-2021 grâce à une aide MRSEI de l'ANR reçue en 2019.

Biblissima+ crée cependant déjà les conditions d'une telle évolution. L'IRHT a déjà mis au point un vocabulaire multilingue (français, italien, espagnol, anglais, arabe) de la codicologie, Codicologia. Biblissima, en agrégeant des ressources étrangères, vise déjà au moins le bilinguisme de ses référentiels. A travers son réseau international, grâce à l'élargissement des projets partenariaux, et grâce à IIIF Collections, Biblissima+ va donner vie à un large bassin de données à rendre interopérables, dont il faudra **extraire les descripteurs multilingues, les aligner, produire des référentiels enrichis, les redistribuer et les rendre collaboratifs, voire participatifs.** La plateforme data.biblissima est prévue pour une ouverture à la science participative. Parmi ses outils, les logiciels d'aide à la traduction du latin et du grec, Collatinus et Eulexis, sont multilingues et fourniront en particulier des noms de personnes à data.biblissima. Il y a donc **dans Biblissima+, à divers niveaux, des mécanismes d'ouverture des référentiels au multilinguisme**. Le cas des descripteurs iconographiques en est un exemple très parlant.

Le bilinguisme est de toute façon un minimum et une nécessité, d'autant plus que Biblissima est très utilisé à l'international. Au-delà, Biblissima+ prépare des briques pour un élargissement international, qui pourrait déjà se traduire par une **convergence de ses référentiels avec le thesaurus très riche du** CERL (Consortium of European Research Libraries), essentiellement issu des métadonnées de fonds d'imprimés anciens. Biblissima a déjà un nombre important d'alignements avec le CERL via la DNB.

Dans tous les projets, les référentiels seront utilisés pour la visualisation des données et les possibilités heuristiques qu'elles ouvrent, en particulier **les SIG et l'analyse de réseaux**, pour lesquels existent déjà d'excellents outils ouverts, y compris chez les partenaires (**EHESS, AOROC** par exemple).

A-II.2/ Fourniture et automatisation des mises à jour des données

Implication : équipe Biblissima+ avec l'ensemble des partenaires

Quelques unes des ressources déjà mises en interopérabilité par le portail Biblissima approchent déjà l'exhaustivité et peuvent être considérées, dans leur première phase, comme des projets achevés, comme *Manuscripta medica* (EPHE-PSL), qui recense tous les manuscrits médiévaux de France à contenu médical, *Pinakes* (IRHT, CNRS), qui recense tous les manuscrits grecs connus, ou *Jonas* (IRHT, CNRS), qui décrit tous les textes en ancien français identifiés et quelques-uns en moyen français. Et pourtant, même eux feront toujours l'objet d'améliorations, d'additions, de mises à jour bibliographiques et de nouveaux projets, et il y aura toujours de nouvelles découvertes. C'est à plus forte raison le cas de ressources encore très ouvertes et en plein développement, comme la plupart des bases de données, bibliothèques numériques et éditions électroniques fédérées par Biblissima. **Ce renouvellement permanent des données pose le problème de leur mise à jour au niveau du portail.** La fourniture automatisée des données et de leur mise à jour était prévue par le guide de mise en interopérabilité du cluster Biblissima, le « Vademecum Biblissima » (https://doc.biblissima.fr/contribuer-a-biblissima : niveau 3 de la mise à disposition des données), mais elle n'a presque pas été mise en œuvre par les partenaires faute de personnel. Elle concerne les ressources du socle Biblissima mises en interopérabilité profonde les unes avec les autres. L'automatisation des mises à jour est essentielle au développement d'une recherche innovante, mais elle suppose au départ des moyens humains chez les partenaires. C'est en effet un énorme défi à relever.

Pour le moment, l'automatisation ne concerne que la fourniture d'un export à jour des bases de données, et ce pour trois ressources : la base *Initiale* de l'IRHT (iconographie des manuscrits des bibliothèques de France), la base *Budé* du CESR (prosopographie des érudits de la Renaissance et de l'époque moderne), la base *Books within Books* de l'EPHE-PSL (fragments de manuscrits hébreux). Si, par exemple, une correction proposée par Biblissima grâce à son travail d'alignement des données n'a pas été répercutée dans la base, l'erreur revient avec le nouvel export et doit être de nouveau corrigée. Il faut donc **consacrer du temps de travail à la répercussion des améliorations dans les bases** en concertation étroite avec les partenaires. Il faut aussi, contractuellement, au nom de l'ouverture des données de l'ESR et de la philosophie générale d'interopérabilité qui permet de valoriser durablement des initiatives isolées et limitées dans le temps, **généraliser les mécanismes de récupération de données structurées, sémantisées et partageables** dans les ressources numériques agrégées le portail.

En effet, un besoin préalable est d'avoir des jeux de données ouverts, accessibles, et à jour, mis à disposition par tous les partenaires et récupérables par l'équipe Biblissima+. Cela procure certes une relative autonomie dans la récupération des données mais reste insatisfaisant et ne permet pas une automatisation des mises à jour à grande échelle. On peut en effet automatiser l'intégration de ressources sans contrôle et sans enrichissement pour atteindre rapidement de gros volumes, mais Biblissima n'a jamais accepté de compromis sur la qualité et la mise en relation des données, et Biblissima+ s'astreint de même à de multiples traitements et vérifications pour garantir la qualité et l'utilité des recherches.

La solution serait probablement de **combiner de façon assez souple plusieurs formats de métadonnées et protocoles d'échange**, permettant à la fois de respecter la nature et structure des données sources et de les moissonner de manière optimale pour les intégrer dans le portail Biblissima+. Une des approches possibles est celle déjà mise en place dans le cadre de Biblissima, qui s'appuie sur **un format XML pivot dont les concepts clés sont dérivés de modèles génériques tels que CIDOC-CRM et FRBRoo.** Dans ce scénario, l'équipe Portail travaille avec les équipes partenaires afin de définir le mapping des données vers ce format XML pivot. Une fois validé ce mapping, les partenaires de Biblissima+ mettent en place un mécanisme de publication automatisée des exports via des web services qui permettent à l'équipe Portail de récupérer périodiquement les jeux de données à jour. A défaut, il est aussi envisageable de s'appuyer sur d'autres formats et mécanismes d'exposition et de moissonnage (entrepôts OAI-PMH, collections IIIF ou API Change Discovery de IIIF pour les bibliothèques numériques, points d'accès SPARQL pour des données RDF etc.). Quelle que soit la solution retenue, les identifiants Biblissima+ (URI) attribués à toutes les entités lors de la phase de traitement et d'alignement seront communiqués aux partenaires qui pourront les stocker dans leurs bases respectives et choisir de les afficher publiquement ou non. **La propagation de ces identifiants URI dans les bases sources facilitera grandement le traitement des données lors des mises à jour.** Ils permettront également aux partenaires de **mettre en place des outils de suivi des mises à jour** des données agrégées dans le Portail, voire de **récupérer le cas échéant les enrichissements** apportés par l'équipe Portail lors de la phase de traitement et d'intégration des données (libellés dans d'autres langues, alignements vers d'autres jeux de données, notes, références bibliographiques, coordonnées géographiques etc.).

Biblissima+ suivra aussi les travaux en cours au sein d'Europeana R&D pour implémenter de nouvelles méthodes d'agrégation et explorera les nouvelles approches et spécifications en matière de partage de

données actuellement en cours de définition au sein de la communauté internationale IIIF (API Change Discovery, API Change Notification).

B. Un arsenal et un écosystème pour de nouvelles découvertes

l'environnement de travail et de service

Une fois présentés ces logiques et mécanismes d'agrégation de données par le portail, c'est-à-dire la cohérence scientifique qui va garantir son efficacité comme infrastructure numérique pour une recherche de pointe sur les cultures écrites anciennes, il convient de présenter le portail lui-même et le bouquet d'outils produits par les pôles de compétences fédérés par le projet, en tenant compte, là encore, de l'expérience de l'équipex Biblissima, des évolutions intervenues depuis huit ans, et de la sociologie des communautés de producteurs et d'utilisateurs.

Insistons d'abord sur un paradoxe lié à cette sociologie des communautés : les projets rivalisent d'innovation technologique, mais les enquêtes auprès des chercheurs montrent que, pour l'immense majorité d'entre eux, ils veulent d'abord d'excellentes numérisations des sources, de la recherche plein texte, de la bibliographie en ligne, si possible grâce à des interfaces très simples d'utilisation. Sur ce point, le grand public les rejoint, ce qui est très important et simplifie la situation d'un certain point de vue. Seule une toute petite frange de la population souhaite disposer d'outils sophistiqués et sait en faire usage. Le pari de Biblissima et plus encore de Biblissima+ est de **mettre les recherches et les technologies complexes au service d'une approche simple et fluide,** et de fournir aux quelques utilisateurs qui le souhaitent des outils avancés qui, un jour, seront peut-être grâce à eux démocratisés. Un exemple peut suffire pour l'expliquer : beaucoup d'internautes rêvent de pouvoir chercher un mot, un nom, directement dans la numérisation d'un manuscrit médiéval, mais pour cela, il faut toute une chaîne invisible d'outils extrêmement avancés de transcription automatique puis de lemmatisation… **L'ergonomie du portail et la conception de l'offre de service de Biblissima+** seront déterminants, d'autant plus si la continuité avec l'offre de service d'autres infrastructures nationales est réussie : celles dont il a déjà été question (Collex-Persée, HAL, Istex, SUDOC, FNE, OPERAS…), mais aussi METOPES pour l'édition et surtout Huma-Num pour l'accompagnement des communautés et l'archivage des résultats. **Biblissima+ mettra à profit et enrichira le contexte exceptionnel que représentent les projets CPER auxquels sont liées ses équipes partenaires** : le centre en humanités numériques du Campus Condorcet (projet Condornum), le programme de Valorisation des patrimoines naturels et culturels auquel participe l'université de Tours (projet VALOPAT), la participation des laboratoires lyonnais à la construction d'un datacenter et d'une plateforme mutualisée pour la recherche, les travaux des équipes de Caen au sein du CPER NumNie (Numérique Normandie), en particulier l'axe 2 - Représentation, enrichissement et interrogation des données et métadonnées et l'axe - 4 Interaction, interfaces homme-machine, accès aux données.

❦

B-I – Cycle de vie des données et outillage de la recherche

Biblissima+ finance des initiatives et des compétences diverses aujourd'hui en partie isolées, et les met en relation. On voit s'organiser ainsi une chaîne d'acteurs et de compétences, à l'échelle nationale mais aussi à l'échelle des sites, qui interviennent dans la production, la mise à disposition, l'enrichissement et la réutilisation de données ouvertes, conformément aux principes FAIR. Ce **chaînage des pôles de compétences s'inscrit dans le cycle de vie ou économie de la donnée scientifique**. Dans le domaine de l'histoire des cultures écrites anciennes, on peut décomposer cette économie de la donnée en six niveaux, depuis la base, les supports physiques, leur conservation et leur numérisation :

6	Publication et dissémination	Outils d'édition : Métopes et MRSH de Caen / Publication : OPERAS / Formation : ILARA etc
5	Visualisation, manipulation et enrichissement	**Pôles de compétences et bouquet d'outils Biblissima+**
4	Accès unifié (portail) et clustering de bases de données	**Portail Biblissima+** / Gallica et éventuelle BibNum de l'ESR / outils de découverte et référentiels
3	Signalement et catalogage	Bibliothèques et archives, programmes scientifiques liés
2	Archivage pérenne et exposition des données	Huma-Num, CINES et BnF
1	Numérisation et conservation des supports physiques	Bibliothèques et archives, musées / IRHT /Collex Persée Istex

- Pour permettre que l'ensemble de l'économie de la donnée sur l'écrit ancien se mette en place, il faut, en tout premier lieu, garantir la qualité des numérisations des documents patrimoniaux et de la bibliographie, qui ont une double fonction de conservation et de transmission, et donc aider les principaux opérateurs (**niveau 1**).
- Le **niveau 2**, dans Biblissima comme dans Biblissima+, est l'affaire des partenaires de l'équipex : l'adoption de formats ouverts et pérennes facilite l'archivage des bases de données auprès de / grâce à Huma-Num et des numérisations à la BnF et au CINES. IIIF 360 encourage l'interopérabilité des bibliothèques numériques.
- Biblissima a contribué, via ses projets fondateurs et ses projets partenariaux, au **niveau 3** (signalement et catalogage).
- Son action principale s'est exercée sur les 3 niveaux suivants. On a expliqué dans la première partie du présent projet les mécanismes d'élargissement du bassin de données de Biblissima+, à partir d'un socle de données acquises par l'équipex Biblissima, et d'automatisation des mises à jour, qui **impliquent les acteurs des niveaux 1-2-3 pour aboutir au niveau 4**.
- Le **niveau 4** favorise la constitution de nouveaux corpus et la naissance de nouvelles problématiques, qui vont être travaillés grâce à l'organisation du **niveau 5** (manipulation et enrichissement des données, visualisation, possibilités de réutilisation) et du **niveau 6**, qui fournit les moyens et structures pour leur diffusion.

A tous les niveaux, il est crucial d'avoir du personnel de niveau IE et IR, mais aussi de l'infrastructure et du matériel. Alors que la partie A de Biblissima+ se concentre surtout sur les premiers niveaux et le niveau 4, la partie B montre comment, pour les communautés travaillant sur les cultures écrites anciennes, on peut fédérer les acteurs et structurer leur action sur le niveau 5 en étroite collaboration avec les acteurs du niveau 6.

Pour agir sur le niveau 5, Biblissima+ propose de donner aux utilisateurs du portail **un accès simple et ergonomique** aux **outils forgés par ses communautés, en proposant une chaîne de traitement des données de la recherche** épousant leur cycle de vie.

Comment faire ? Là encore, intervient la sociologie des communautés et des usagers. Les environnements de travail très sophistiqués sont conçus par des communautés en fonction de leurs propres besoins, et ne conviennent en général qu'à un petit nombre d'usagers. Un environnement de travail trop riche, débordant de possibilités, trouble l'usager plus qu'il ne l'aide. Les recensements d'outils se multiplient, mais il ne semble pas qu'ils atteignent vraiment des cibles très nombreuses. Et pourtant, il faut faciliter l'accès à toute la palette des outils, et aider les équipes très innovantes.

Sur le même modèle que celui des données, qui a fait ses preuves dans la réalisation de l'équipex Biblissima, Biblissima+ propose de **fédérer (et donc financer) des initiatives** qui se développent en dehors du portail, pour leur propre compte, **qui gardent leur indépendance, mais trouvent grâce au portail de nouveaux bassins de données et d'utilisateurs – et donc de nouveaux défis**, et dans l'infrastructure et ses communautés **un lieu de réflexion et de confrontation intellectuelle**. Pour éviter que les outils ne se développent de façon totalement anarchique, non concertée, sans trouver de nouveaux utilisateurs, nous proposons **une organisation des outils selon le cycle de vie des données, qui dessine un chaînage des outils**, besoin largement reconnu aujourd'hui, dont témoigne par exemple le projet DAHN (« Dispositif de soutien à l'archivistique et aux humanités numériques », DGRI, plan SHS 2019). Cette organisation dessinera de grands ensembles (par ex. 3D, HTR, édition TEI…) qui **rassembleront les acteurs du domaine** : dans ces groupes se confronteront les idées, les solutions, se détermineront de grandes orientations, et il devrait en résulter une fécondation

réciproque et, qui sait, une simplification du paysage au fil des ans. **En retour, Biblissima+ récupère les métadonnées référençables et les annotations et livrables** que les chercheurs sont prêts à partager, de façon à enrichir le portail et ses référentiels en données validées issues des projets financés.

Comme l'a fait Biblissima pour les projets de recherche, Biblissima+ financera des développements à deux niveaux, des briques principales financées dès le départ par le projet et dans la durée, et des briques supplémentaires fédérées au fil des appels à projets annuels, pour des développements techniques à ce jour imprévisibles.

Biblissima+ jouera ainsi son **rôle dans la structuration et l'enrichissement du centre en humanités numériques du Campus Condorcet, décrit par le projet CPER Condornum**, dont il partage les principes : favoriser les outils généricisables et leur adoption, éviter les redondances avec d'autres projets, favoriser la réutilisation des résultats et des outils. **Par sa structure multipolaire, il fera de même au niveau national, en permettant aux équipes de concilier les logiques de site** (également articulées localement avec des CPER) **et cette logique nationale**. Il développera ou contribuera au développement de trois niveaux d'outillage de la recherche sur les cultures écrites :

- les **nouveaux outils** informatiques pour l'analyse des artefacts anciens et des textes qu'ils portent, qui vont faire prendre à la recherche **un vrai tournant numérique et susciter de nouvelles découvertes** ; les outils devenus classiques, dont les chercheurs ont déjà l'habitude, ne seront pas mentionnés ici, car on considère qu'ils font déjà partie de l'environnement de travail des communautés ;
- les installations physiques d'acquisition et de traitement des données qui vont permettre de réaliser les ambitions (équipement de numérisation, fixe et mobile ; équipement et temps de travail pour l'analyse des matériaux ; puissance de calcul pour l'IA) ;
- le service, en ligne et en présentiel, pour l'accompagnement des utilisateurs, leur montée en compétence et en autonomie, la réponse aux nouveaux besoins.

Dans cette partie B-I, nous présentons d'abord les pôles de compétences de Biblissima+ avec l'indication succincte des environnements de travail qu'ils proposent – qui sont déjà en soi des outils avancés –, puis les principaux regroupements d'outils en suivant le cycle de vie des données de la recherche (p. 32 à 51).

… une infrastructure multipolaire …

Biblissima+ est construit sur **un socle composé des partenaires de l'équipex Biblissima (2012-2020)** : porté par l'EP Campus Condorcet, Biblissima a associé de grands acteurs français de la recherche et de la conservation sur le patrimoine écrit du Moyen Âge et de la Renaissance : la BnF, le CESR (Tours), le CIHAM (Lyon-Avignon), le CJM (ENC-PSL), SAPRAT (EPHE-PSL), l'IRHT (Paris-Orléans), le CRAHAM et la MRSH de Caen, et, à partir de 2017, les Archives nationales.

Biblissima a financé d'importants programmes de numérisation et de signalement à la BnF, participé à la création de Reliures.bnf et à la refonte de Mandragore, favorisé le passage de Gallica à IIIF et fourni ses référentiels à data.bnf. Ces évolutions expliquent que la BnF, qui ouvre largement ses ressources, n'ait plus besoin de faire partie du consortium, en revanche **une large ouverture aux services d'archives** est désormais nécessaire, ce qui se réalise grâce à la participation du Service Interministériel des Archives de France pour les Archives nationales (AN) et le portail FranceArchives. L'élargissement du périmètre scientifique de Biblissima+ lui permet de **faire travailler ensemble des communautés qui ne l'avaient pas fait jusqu'ici**. AOROC, qui voit déjà travailler ensemble archéologues et spécialistes des traditions textuelles, regroupe **toutes les spécialités en humanités numériques**. **Le volet épigraphique qui manquait à Biblissima est au complet** avec HiSoMA et AOROC pour l'Antiquité, le CESCM pour l'époque médiévale, le CESR pour la Renaissance et le CJM pour l'époque moderne. Le CRH et les Editions de l'EHESS apportent une riche contribution sur les textes antiques et médiévaux, l'iconographie médiévale, la TEI et les référentiels, qui manquait au précédent équipex. Désormais, un **pôle national d'éditions savantes en TEI d'une richesse exceptionnelle** se constitue grâce à Biblissima+, et s'ouvre à la musique, apportant **la dimension sonore** qui manquait. Les recherches pionnières sur la **reconnaissance de caractères manuscrits dans toutes les langues et systèmes graphiques**, à l'aide de **l'intelligence artificielle** et en lien avec les autres clusters de Biblissima+ (la numérisation IIIF, la TEI, la lemmatisation, la texto- et stylométrie) s'enrichissent de l'intégration non seulement de Scripta-PSL mais aussi de l'entreprise TEKLIA, qui travaille étroitement avec l'IRHT. Avec l'interaction du CRC et des équipes de philologues, historiens, humanistes numériques, **l'analyse physico-chimique des matériaux est enfin intégrée à l'écosystème**, le CRC s'engageant dans une

démarche novatrice de mise à disposition et de cherchabilité de ce type de données.

Le partenariat Biblissima+ n'est pas refermé sur lui-même, **l'infrastructure portant en elle-même les mécanismes d'une ouverture maîtrisée** : à d'autres équipes nationales et internationales via les appels à projets Biblissima+ ; aux autres grandes entreprises d'envergure nationale et internationale (IR, TGIR, équipex+, ERIC…) par la mise à disposition de ses référentiels et la mise en place de mécanismes d'intégration et d'échange.

Biblissima+ est donc une infrastructure multipolaire reposant sur des pôles de compétences répartis sur l'ensemble du territoire national, qui porte en soi les mécanismes d'une expansion raisonnée.

Sur le Campus Condorcet, l'équipe portail Biblissima au cœur du dispositif

L'équipe portail Biblissima fait partie du dispositif du Campus Condorcet, où elle travaille en lien direct avec les équipes du GED, faisant le lien entre toutes les communautés ESR, Culture, recherche et établissements de conservation. Elle assure la convergence des référentiels et leur réutilisation par les équipes, en particulier Persée, HAL, Istex, Huma-Num et OpenEdition / OPERAS, la mise en interopérabilité des ressources numériques et la pertinence des systèmes de récupération et mise à jour des données. Elle développe et maintient les sites centraux de Biblissima+ en s'adaptant aux besoins des utilisateurs et en assurant sur toute la chaîne de traitement le respect des principes FAIR. Elle assure l'acculturation des communautés et leur accompagnement, en particulier sur tous les développements du protocole IIIF et son alliance avec Omeka / Omeka S. Avec le responsable scientifique et technique, elle gère les appels à projets. Les emplois sont pérennisés par le Campus Condorcet.

Actuellement : https://biblissima.fr.
Environnement proposé : un portail simplifié et un environnement d'outils, à construire.

À Aubervilliers, Paris et Pierrefitte, un « couteau suisse » : les acteurs du Campus Condorcet, des AN et de PSL

Chacun des acteurs de Paris et de la région Île-de-France a développé des compétences multiples, avec de nombreuses collaborations qui s'entrecroisent. Le plus simple est, comme pour les laboratoires en région, de présenter chacun d'entre eux, selon une logique institutionnelle, le bouquet Biblissima+ étant destiné, ensuite, à mettre en valeur les regroupements cohérents.

☞ Les AN et FranceArchives

Les Archives nationales (AN), à Paris et Pierrefitte, assurent la collecte, le traitement, la conservation et la mise en valeur des archives centrales de l'État. Les deux enjeux du numérique et de la transmission des savoirs sont centraux pour les AN. Le premier a donné lieu à une refonte du système informatique et à la création d'un référentiel des producteurs d'archives ; le second vise à encourager et à faciliter la transmission des savoirs et des savoir-faire développés aux AN, en particulier les compétences en matière de sciences auxiliaires de l'histoire (paléographie, diplomatique, codicologie, sigillographie, onomastique), mais aussi l'expertise technique mobilisée pour la restauration et la numérisation des documents et l'archivage électronique.

Piloté par le **Service interministériel des Archives de France (direction générale des patrimoines, ministère de la Culture)**, FranceArchives est un point d'entrée unique offrant un accès fédéré aux millions de documents d'archives de 75 services d'archives partenaires (dont les 3 SCN des Archives nationales, le Service historique de la Défense et les Archives diplomatiques, la Médiathèque de l'architecture et du patrimoine, 48 services d'Archives départementales, 10 services d'Archives municipales ou métropolitaines, 9 établissements publics et 2 associations) et, à terme, de tous les services d'archives français. Le portail intègre 52 257 inventaires et reçoit 262 000 visiteurs uniques par mois en moyenne (en 2020). Les jeux de données qu'il diffuse sont accessibles en open data (https://francearchives.fr/fr/open_data). Il est l'agrégateur national vers le Portail européen des archives.

Portail FranceArchives : (https://francearchives.fr/fr/)

☞ Le CJM et le CTHS (ENC-PSL)

Le **Centre Jean-Mabillon (CJM)** est le laboratoire de recherche de l'École nationale des chartes (ENC-PSL). Le CJM incarne la dimension collective, interdisciplinaire et internationale des activités de recherche dans le domaine des sciences historiques et philologiques telles qu'elles sont enseignées à l'ENC. Il occupe une position singulière dans la recherche par ses relations étroites avec les institutions de conservation du patrimoine écrit, en particulier grâce à la présence active en son sein de conservateurs du patrimoine et des bibliothèques. Les modalités de la fabrique du patrimoine, objet et source d'histoire tout à la fois, sont ainsi placées au cœur des réflexions et des travaux du Centre. Cette complémentarité contribue de même à placer le CJM au carrefour des débats autour des nouvelles modalités d'édition des documents : l'interrogation sur les formes et les usages du texte et de l'image à l'âge du numérique intéresse tant les spécialistes des humanités (histoire, philologie, littérature, art, etc.) que ceux des technologies de l'information et de la communication, mais aussi les pouvoirs publics confrontés à des défis de masse tels que la diffusion des collections patrimoniales et l'exploitation de données publiques. Le CJM, avec la Mission projets numériques de l'ENC,

travaille sur l'OCR des textes anciens, leur édition en TEI, la fouille de corpus, la texto-et la stylométrie, la lemmatisation, le développement de référentiels.

L'ENC développe en particulier des référentiels prosopographique, géographique et institutionnel avec le **Comité des travaux historiques et scientifiques** (**CTHS**), qui lui est rattaché. Le CTHS se trouve à la tête d'un réseau de plus de 3 000 sociétés savantes et fédère des scientifiques membres de prestigieuses institutions, des érudits locaux, de jeunes chercheurs et joue ainsi un rôle fondamental dans la construction et la transmission des savoirs. Deux grandes bases de données y sont développées : *La France savante (XVII^e^-XX^e^ siècle)* et la réédition numérique du *Dictionnaire topographique de la France*, commencé au XIX^e^ siècle (plus d'un million d'entrées).

Environnements proposés : « Chaîne de traitement numérique des manuscrits latins et romans », « Chaîne de traitement des cartulaires et des documents d'archives ».

☞ Le CRC au MNHN

Le **Centre de recherche sur la conservation (CRC)** regroupe des chimistes, physiciens, biologistes, historiens de l'art ou historiens des sciences et des techniques. Leurs travaux s'intéressent aux objets de muséums, de musées, d'archives et de bibliothèques mais également aux monuments historiques, et visent en particulier la connaissance des matériaux du patrimoine et l'étude de leurs processus d'altération, ainsi que le développement de méthodes de conservation préventive et curative. Dans le cadre du projet REMAC - A la REcherche des MAnuscrits de Chartres, financé en 2017 par la Fondation des Sciences du Patrimoine (LabEx PATRIMA), le CRC a travaillé avec la médiathèque de Chartres, l'IRHT, le laboratoire Dynamiques Patrimoniales et Culturelles et le Laboratoire d'Optique et Biosciences pour développer des techniques d'imagerie scientifique et de traitement de l'image qui permettent aux historiens de lire et d'étudier à nouveau les manuscrits chartrains. Un deuxième axe consiste à mettre au point de nouvelles techniques d'imagerie et de microscopie pour visualiser l'état de dégradation des parchemins. Actuellement, le CRC travaille avec la BM d'Avranches et le CRAHAM sur les matériaux des manuscrits du Mont-Saint-Michel et prépare un projet sur les bibliothèques médiévales de Normandie prouvant les convergences entre sciences des matériaux et histoire des bibliothèques, et les découvertes qu'elles rendent possibles.

☞ L'EHESS et le CRH

L'École des hautes études en sciences sociales produit et transmet des savoirs sur les sociétés humaines en organisant un dialogue interdisciplinaire permanent entre histoire, sociologie, anthropologie, économie, philosophie, géographie, études littéraires, psychologie et sciences cognitives. Le **Centre de recherches historiques (CRH)** est l'une de ses 40 unités. Ses chercheurs, qui travaillent sur des aires culturelles et des périodes très variées (du Moyen Âge à nos jours), ont en commun le souci de la longue durée, la réflexion sur l'histoire sociale, la pratique de l'interdisciplinarité, le jeu d'échelles entre le local et le global, le renouvellement des sources de l'histoire, ainsi que l'ouverture internationale des objets, des programmes et des pratiques de recherche. Antiquisants et médiévistes ont développé plusieurs bases de données (Images, Thema, Quodlibase) et environnements de travail (ALGreM, Savoirs) faisant dialoguer TEI, SIG, ressources bibliographiques, alimentant le nouvel entrepôt de données Didomena et le référentiel DATU de l'EHESS, récemment publié. Les **Editions de l'EHESS** abritent une cellule éditoriale multisupport qui sera basée sur le Campus Condorcet à partir de 2021, qui travaille en étroite collaboration avec la MRSH et les Presses universitaires de Caen, autour de la chaîne éditoriale Métopes.

Environnements proposés : ALGreM, Anthologie du livre grec manuscrit, et « Savoirs », environnement de lecture enrichie de ressources bibliographiques, en écho direct à la partie A de Biblissima+.

☞ e-Scripta à l'Université PSL

Au sein de l'Université PSL, le programme ***Scripta-PSL. Histoire et pratiques de l'écrit*** vise à intégrer les sciences fondamentales de l'écrit (paléographie, codicologie, épigraphie, histoire du livre, etc.), avec d'autres SHS (linguistique, histoire, anthropologie, etc.) et les humanités numériques et computationnelles, en intégrant pleinement l'intelligence artificielle (IA) dans ses pratiques de recherche. Le programme est porté par l'EPHE-PSL et l'EFEO, en association avec l'ENS-PSL, l'ENC-PSL, le Collège de France et l'IRHT, en partenariat avec l'EHESS. La recherche se décline en six axes : « Écritures et langues » ; « "Pages" – champs visuels pour la lecture » ; « Écritures exposées – inscriptions de l'espace » ; « Pratiques documentaires, anciennes et modernes » ; « Circulation de l'écrit – processus de canonisation » ; « Défis pour l'édition scientifique numérique ». Le pôle numérique de Scripta, **e-Scripta**, s'est attelé à la création d'un environnement de travail, eScriptorium, qui combine des outils pour la transcription automatique et l'annotation approfondie de textes et d'images (paléographique, philologique, historique et linguistique), conçu dès le départ pour une grande variété d'écritures et de langues, surtout les langues rares et historiques, couvrant plus de 3 millénaires, et presque le monde entier.

Environnement proposé : le programme en soi est un projet d'environnement, combinant Kraken (HTR), Archetype (outil d'annotation et d'analyse des images) et eScriptorium, l'environnement de travail proprement dit.

☞ L'ENS-PSL, l'EPHE-PSL et les programmes d'AOROC et de SAPRAT

L'École Pratique des Hautes Etudes (EPHE-PSL) a été créée pour promouvoir des méthodes de formation à la recherche par la recherche, dans le cadre de ses laboratoires et de ses séminaires. Le champ couvert par ses formations et ses nombreuses équipes de recherche inclut des disciplines aussi variées que l'histoire, la philologie, l'archéologie, l'histoire de l'art et la linguistique, l'étude des religions abordée sous tous les angles disciplinaires, mais aussi les sciences de la vie et de la terre et les sciences cognitives. L'EPHE développe depuis plusieurs années d'importants programmes utilisant le numérique, grâce à des collaborations entre ses sections scientifiques et avec des laboratoires d'informatique (l'INRIA en particulier). En 2020, elle crée un Institut des langues rares (**ILARA**) sur le Campus Condorcet, qui, avec la collaboration du LLACAN et du LACITO, développera des logiciels de transcription automatisée d'enregistrements oraux utilisant l'intelligence artificielle.

Co-tutelle avec le CNRS et l'ENS-PSL du laboratoire **Archéologie et philologie d'Orient et d'Occident (AOROC)**, l'EPHE-PSL vient d'y concentrer plusieurs de ses chercheurs les plus impliqués dans le numérique et dans e-Scripta, qui y rejoignent des spécialistes de SIG et de reconstitution 3D.

Le laboratoire **Savoirs et pratiques du Moyen Âge au XIX^{e} siècle (Saprat)**, partenaire de Biblissima, voit quant à lui le développement de deux programmes importants pour Biblissima+ : **Multipal**, fruit du travail du Groupe de recherche transversale en paléographie (GRTP : EPHE, ENC, Collège de France), et **e-Signa, portail de l'emblématique médiévale**, qui mutualise les données (armoiries, cimiers, devises, signatures) issues des bases de données et des corpus de sources médiévales recensant ces signes et leurs supports : sceaux, monuments, vitraux, manuscrits, etc. (DEVISE, SIGILLA, ARMMA, COLLECTA, RCPPM, BIBALE, Corpus Vitrearum, etc.).

> Environnements proposés : Multipal, album paléographique et tutoriel interactif multilingue ; e-Signa, interface d'identification des signes et emblèmes et SIGISCRIPT, outil de traitement de l'épigraphie (encodage TEI et 3D, appliqué au sceau).

☞ Le GED

Plus de 50 bibliothèques, fonds documentaires et services d'archives, actuellement dispersés sur 25 sites en Île-de-France (dont les fonds de l'EHESS, de l'EPHE, de l'IRHT), apportent leurs collections au **Grand équipement documentaire du Campus Condorcet (GED)**, soit un million de documents : principalement des livres et des revues, mais également des archives scientifiques, des photographies, des films, des cartes, des enregistrements sonores, sous forme physique ou numérique. S'appuyant sur les expériences, notamment numériques, les plus innovantes, le GED est conçu comme un laboratoire partagé pour la recherche en sciences humaines et sociales. Il est le point de convergence entre les disciplines, les étudiants et les chercheurs. Il proposera un ensemble de services au cœur desquels se trouve l'accès libre à une documentation matérielle et dématérialisée, ainsi qu'une offre appuyée sur les humanités numériques. Dans Biblissima+, le GED, porteur du projet de CPER Condornum déposé fin 2019, joue un rôle essentiel par son positionnement au carrefour de toutes les communautés impliquées, sur le Campus, mais aussi au niveau national par son rôle dans le GIS Collex-Persée et ses liens avec les grandes entreprises d'IST et les réseaux nationaux et internationaux de bibliothèques et archives.

> Environnement proposé : Système d'information archives, Système de gestion de bibliothèque Alma et outil de découverte Primo, bibliothèque numérique Omeka-S, pilotage de numérisation NumaHop, Condornum.

☞ L'IRHT, à Aubervilliers, Paris et Orléans

Unité propre du CNRS, l'**Institut de recherche et d'histoire des textes (IRHT)** est implanté sur le Campus Condorcet, à Orléans et à Paris. Le laboratoire est un centre de compétences de renommée internationale. Il pratique la recherche fondamentale sur les documents manuscrits médiévaux (sur papyrus, parchemin, papier, *ostraka*…) et sur les textes écrits sur ces supports dans les principales langues de culture du pourtour méditerranéen, de l'Antiquité à la Renaissance. L'examen des fonds documentaires conduit à la production de répertoires et de catalogues. La recherche privilégie le travail de datation des manuscrits, l'édition critique des textes, l'étude de la transmission des œuvres et des échanges culturels qui l'accompagnent. D'importants programmes de recherche sont pilotés à l'IRHT : bibliothèques médiévales de France, droit musulman pré-moderne, bibliothèques byzantines, œuvres à succès de la littérature romane médiévale, pratique du commentaire des Sentences de Pierre Lombard dans les universités européennes, traductions latines d'œuvres vernaculaires, paléographie numérique, etc. Ils font appel à de nombreux partenariats français et internationaux. Depuis toujours pionnier dans l'usage des nouvelles technologies et chargé de la reproduction des manuscrits des BM et BU de France pour les ministères de la Culture et de l'Enseignement supérieur, l'IRHT a pris très tôt le tournant du numérique et développé des bases de données et des outils novateurs. Il est aujourd'hui très engagé dans le *machine learning* pour la reconnaissance automatisée de formes et de caractères et dans l'édition électronique.

> Environnements proposés : Telma-ANACLET, chaîne d'édition électronique de corpus larges et non-enrichis, offrant des outils de traitement des textes a posteriori ; pôle d'édition électronique de sources.

À Caen, un laboratoire d'édition et d'annotation de sources : CRAHAM et MRSH-PDN

Le **Centre de Recherches Archéologiques et Historiques Anciennes et Médiévales. Centre Michel de Boüard (CRAHAM)** dispose de savoir-faire établis dans plusieurs domaines : étude des sources textuelles anciennes (manuscrits, sources diplomatiques, textes hagiographiques ; historiographie) ; archéologie de terrain (prospection, diagnostic et fouille, rapports) et de laboratoire (archéométrie-céramologie : analyses chimiques de céramiques archéologiques et d'argiles) ; paléoanthropologie ; numismatique ; archéomatique (traitement informatique des données, SIG). Il est engagé avec le CRC dans un grand programme sur les sciences des matériaux appliquées à l'histoire des bibliothèques médiévales de Normandie.

Le **CRAHAM** est fortement impliqué dans le domaine des humanités numériques **en étroite collaboration avec le Pôle Document numérique (PDN) de la MRSH de Caen** et les Presses universitaires de Caen : revue électronique *Tabularia* ; bases de données Nummus et ITAM (monnaies) ; base de données SCRIPTA (chartes normandes X^{e}-XIII^{e} siècle) ; l'outil *e*-Cartae (*outil d'édition critique et de publication multimodale (papier et numérique) des chartes médiévales)*, programmes d'éditions de textes papier et électronique dans les collections des PUC *Fontes et paginae*, *e*-Cartae et *Thecae* (RIN "Norécrit" pour les sources médiévales de la Normandie, ANR Actépi, programme "Ichtya" pour les sources ichtyologiques) ; programme de « Bibliothèque virtuelle du Mont Saint-Michel » consacré à l'étude et la valorisation des fonds patrimoniaux conservés à Avranches, RIN "Cornum" (Contenus et Corpus numérique), d'autorités partagées en XML-TEI (*thesauri* lieux, personnes, œuvres, pièces liturgiques), constitution de bases prosopographiques (projet Vexicaen, RIN Normonde). Ces programmes répondent aux principes *FAIR*, plusieurs d'entre eux sont menés en collaboration avec les *Consortia* CAHIER, COSME et MASA d'Huma-Num et certains ont été lancés ou soutenus par Biblissima (Thesauri ; BVMSM ; Thecae). Le CRAHAM et le PDN font valoir dans Biblissima+ leur expérience inégalée en édition de sources anciennes, qui offre la spécificité de couvrir la chaîne complète d'édition, depuis le fichier du chercheur jusqu'à la publication et la diffusion par une maison d'édition, les PUC, en collaboration étroite avec l'IR METOPES (la « chaîne de Caen » utilisée par les autres partenaires).

Le **PDN**, spécialisé dans le domaine des humanités numériques, conçoit, développe et met en œuvre des outils numériques et des méthodes de travail pour les programmes de recherche en humanités et sciences humaines et sociales (SHS) avec une approche centrée sur les données. Il est fortement impliqué dans les consortia CAHIER et MASA de la TGIR HumaNum.

Environnement proposé : laboratoire d'édition et d'annotation de sources.

À Lyon et Avignon, archéologie, épigraphie et science des textes : le CIHAM et HiSoMA

A Lyon, deux laboratoires seront impliqués dans Biblissima+. Ils associent des antiquisants, des tardo-antiquisants et des médiévistes travaillant dans des domaines multilingues du pourtour méditerranéen et associant étroitement science des textes et archéologie, avec un très fort intérêt pour l'épigraphie et une grande implication dans les humanités numériques et le consortium TEI.

HiSoMA (Histoire et sources des mondes antiques), du fait de sa composante « Sources chrétiennes » porteuse du projet Biblindex, de ses programmes scientifiques regroupés dans l'axe « Edition, archives, humanités numériques », de son programme d'éditions épigraphiques et de ses missions archéologiques (Chypre, Egypte, désert oriental), de son implication dans l'ERC Synergy DHARMA etc., représente pour Biblissima+ un apport précieux et une ouverture réelle. L'un des atouts de ses équipes est de pouvoir collaborer, d'une part, avec le pôle ingénierie scientifique du laboratoire et, d'autre part, avec le PSIR (Pôle Système d'Information et Réseaux) de la Fédération Maison de l'Orient et de la Méditerranée dont HiSoMA est une unité constitutive. Elle a tiré parti des méthodes et des ressources diffusées par diverses communautés scientifiques, tant dans la région Auvergne-Rhône-Alpes qu'en France et à l'international (Humanistica – association francophone des humanités numériques, consortium Cahier – Corpus d'auteurs pour les humanités : informatisation, édition, recherche, consortium MASA, IR Métopes, projet de la Text Encoding Initiative pour le balisage et l'encodage des textes en xml, les outils d'EpiDoc, etc.).

Environnements proposés : éditions épigraphiques en TEI ; le site Biblindex.

Le champ d'activité du **CIHAM (Histoire, Archéologie, Littératures des mondes chrétiens et musulmans médiévaux, Lyon-Avignon)** couvre le millénaire médiéval, avec une spécialisation sur les trois derniers siècles du Moyen Age. Géographiquement, les enquêtes portées par le CIHAM concernent la vallée du Rhône et le quart sud-est de la France, l'Italie, la péninsule Ibérique et le Maghreb avec des extensions vers le nord de la France et le Proche-Orient. Très variées, les sources mobilisées proviennent des archives, des bibliothèques, des musées mais aussi des archives des chercheurs. Elles sont aussi le produit de trois chantiers archéologiques majeurs (Albalat en Espagne, L'Isle-sur-la-Sorgue en Vaucluse, et La Couronne dans l'Allier). Dans la configuration de l'UMR pour les cinq années à venir (contrat 2021-2025), un axe transversal « Humanités numériques » est maintenu, reliant les différents thèmes scientifiques et nourri par leurs diverses activités. Cet axe qui s'appelle désormais Corpor@Tech s'appuie sur

l'expérience accumulée depuis plusieurs années grâce aux travaux de Marjorie Burghart et est relancé grâce au recrutement d'une IE en 2018. Le CIHAM s'implique résolument dans la science ouverte.

> Environnements proposés : édition électronique et formation à l'édition électronique des textes latins, grecs et arabes.

À Poitiers, l'épigraphie médiévale : le CESCM

Depuis 1974, le *Corpus des inscriptions de la France médiévale* (*CIFM*) du **Centre d'études supérieur de civilisation médiévale (CESCM)** recense, publie et étudie les inscriptions du Moyen Âge tracées sur bois, sur pierre ou sur métal et conservées dans les églises ou autres bâtiments. Les trois quarts de la France sont désormais couverts (66 départements, 8 régions). L'Est, le Nord et la région parisienne seront les grands chantiers à venir. Ces quarante ans d'édition s'accompagnent d'autant d'années d'études d'épigraphie médiévale, d'enseignement et de développement de la discipline, qui permettent aujourd'hui de réviser les anciens tomes et de proposer des Hors-Séries thématiques en complément de la collection habituelle. Les anciens volumes sont accessibles sur Persée. TITULUS est un projet d'épigraphie médiévale numérique du CESCM. Son but est de proposer une nouvelle diffusion accessible à tous des sources épigraphiques du 8e au 15e siècle du territoire français et de nouveaux moyens d'exploitation de ces textes, à travers une édition électronique des nouveaux volumes du *CIFM* et une base de données.

> Environnement proposé : atelier d'épigraphie médiévale, de l'acquisition de nouvelles sources, la prise d'images 3D, à l'analyse automatisée d'images et la réflexion sur les matériaux et les techniques.

À Tours, l'étude du patrimoine écrit et musical de la Renaissance : CESR

Les **BVH** (Bibliothèques Virtuelles Humanistes) sont un programme de recherche du CESR portant sur le patrimoine écrit, manuscrit et imprimé, de la Renaissance. Les actions scientifiques concernant les données bio-bibliographiques, l'édition numérique d'archives manuscrites et d'imprimés, les matériaux typographiques, sont développées systématiquement avec des acteurs présents dans un vaste territoire qui va du bassin ligérien (Orléans, Tours, Bourges, Châteauroux, Le Mans, Angers, etc.) aux frontières de la Nouvelle Aquitaine (Bordeaux). Il s'agit souvent de bibliothèques relevant de la tutelle du MCC et du MESRI, de laboratoires de recherche en informatique (LIFAT), analyse des matériaux (IRAMAT), humanités numériques (MSH Val de Loire). A terme, les BVH s'engagent à procéder à un maillage plus serré du « grand Centre-Ouest ». Les BVH coordonnent également, pour le Ministère de la Culture, la publication des *Catalogues Régionaux des Incunables* aussi bien dans la version papier que numérique.

L'ensemble des corpus et les données qui les composent (images, textes, personnes, lieux, institutions, etc.) alimentent d'ores et déjà Gallica, Europeana, EDIT16 et sont alignés avec les principaux *repositories* et catalogues internationaux (VIAF, USTC, ISTC). Ils font également l'objet de recherche dans le domaine de la linguistique, des OCR, de l'étude des provenances. Les données prosopographiques et géographiques concernant les incunables (CRIICO) et les auteurs français du 16e siècle (*Bibliotheques françoises* de La Croix du Maine et de Du Verdier) seront enrichies, dans les années à venir, par celles développées au sein de projets en cours, portant sur le corpus poétiques manuscrits (projet POMONE et manuscrits italiens 15e-17e siècles), sur le dépouillement et la transcription d'archives locales (projet RENUMAR) et d'archives de chercheurs (« fonds Pierre Aquilon » sur les inventaires des bibliothèques privées), sur les provenances des imprimés en italien dans l'espace francophone (projet EDITEF), pourront alimenter le portail Biblissima+ et faire l'objet de dépôts auprès de partenaires internationaux (ICCU, CERL, IdREF, DARIAH, etc.).

Au sein du CESR, l'équipe **Ricercar** introduit dans Biblissima+ la chaîne de traitement de la donnée musicale. Le programme articule ses activités autour de 3 pôles : 1. La recherche en musicologie ; 2. La valorisation du patrimoine musical ; 3. La constitution d'un fonds documentaire. Les actions menées au sein du programme sont multiples et se concrétisent sous des formats différents en fonction des orientations proposées par son directeur et son équipe. Elles reflètent également la dimension internationale dans laquelle s'inscrit le programme de recherche, en participant régulièrement aux colloques et congrès, en organisant des séminaires, des colloques et des journées d'études sollicitant la communauté internationale ; en accueillant des longs séjours pour des professeurs invités (Le Studium). Outre la mise à disposition de ressources – sources imprimées et manuscrites, documents iconographiques et/ou sonores, ou bien des transcriptions en notation moderne –, Ricercar développe des outils numériques pour répondre au besoin des projets dont il est le porteur ou le partenaire. Il intègre au fur et à mesure les avancées technologiques du MEI (Music Encoding Initiative) en étroite collaboration avec ses institutions de tutelles et de la TGIR Huma-num. C'est au travers d'un dispositif de médiation culturelle scientifique, le *Cubiculum musicae*, que l'équipe propose un équipement d'immersion musicale et visuelle, élaboré selon des critères scientifiques afin de replacer une œuvre musicale dans son contexte. Enfin, les projets de recherche sont valorisés à travers des publications web de corpus de musique des 15e et 16e siècles et les travaux scientifiques à travers la collection « *Épitome musical* » (Brepols Publishers) qui compte aujourd'hui près de 100 volumes (monographies, actes de colloque, études, éditions critiques) ; — ou encore au

travers de la collection « *Musicologie* » (Classiques Garnier) qui regroupe désormais une dizaine d'ouvrages.

Environnements proposés : Édition numérique TEI, bases de données et corpus de documents de la Renaissance, formations TEI ; un Hub entre IIIF (l'image de la partition) – MEI (son encodage XML analogue à la TEI pour les textes) – l'audio (synthétique ou non).

… sept clusters pour un chaînage des outils …

Une fois les corpus constitués, via les campagnes de numérisation ciblées, comme les a encouragées Biblissima, ou via les recherches spécifiques menées sur le portail et produisant de nouveaux sets de données, viennent les étapes de traitement et d'enrichissement. Indépendamment du grand travail collaboratif autour des référentiels, présenté en partie A, et indépendamment des recherches qu'ils permettront – par ex. des visualisations cartographiques (SIG), ou de l'analyse de réseaux –, les recherches actuelles des pôles de compétences Biblissima+ en région et à Paris ont fait émerger **sept domaines à la pointe de l'innovation, avec leurs problématiques, leurs outils et leurs formations, organisés en chaîne de traitement des données de la recherche** épousant leur cycle de vie, de l'acquisition de numérisations interopérables, y compris en 3D, jusqu'à l'étude approfondie des textes :

- l'acquisition de corpus d'images interopérables (imagerie 2D, 3D, hyperspectrale), soutenue par le service IIIF 360 opéré par Biblissima+, le Campus Condorcet et la TGIR Huma-Num ;
- la cherchabilité des données d'analyse physico-chimique des matériaux des artefacts anciens ;
- l'intelligence artificielle au service de la lecture automatique d'écritures anciennes et de la reconnaissance de formes, avec intégration à MésoPSL d'un supercalculateur dédié ;
- une plateforme d'analyse et apprentissage des divers systèmes graphiques anciens quels qu'ils soient ;
- un pôle exceptionnel consacré à l'édition savante en TEI ;
- le repérage, l'édition et l'intégration des sources sonores anciennes dans le portail ;
- les outils les plus révolutionnaires d'analyse automatisée des textes.

Biblissima+ crée donc **7 clusters** qui rassembleront les acteurs du domaine : dans ces groupes se confronteront les idées, les solutions, se détermineront de grandes orientations. Cette **fabrique du consensus** créera **une fécondation réciproque mais aussi une simplification du paysage au fil des ans, et permettra l'émergence de nouvelles idées**, qui feront évoluer les clusters.

En retour, Biblissima+ récupérera les métadonnées référençables et les annotations et livrables que les chercheurs seront prêts à partager, de façon à enrichir le portail et ses référentiels en données validées issues des projets financés. Le portail Biblissima+ donnera un accès convivial et simplifié aux outils, de façon que l'utilisateur puisse enrichir et diffuser les jeux de données constitués grâce à l'interopérabilité des ressources. **Ainsi sera mis en œuvre le cycle de vie des données, par une infrastructure numérique multipolaire de recherche fondamentale et de service.** Ces sept clusters de pointe rassemblant plusieurs pôles, tous financés, ayant intérêt à **mutualiser leurs compétences et à réfléchir ensemble** sur leurs propres pratiques, ont une gouvernance **totalement décentralisée**. Si l'infrastructure numérique est coordonnée par l'équipe portail Biblissima+, exerçant sur le Campus Condorcet, **les clusters ont leur vie propre, tout en étant reliés à l'ensemble et entre eux par divers moyens** :

- Chaque cluster a un coordinateur, qui échange régulièrement avec la responsable scientifique et technique et l'équipe portail Biblissima+, et rend compte deux fois par an au comité de direction de Biblissima+ de ses activités.
- Chaque cluster a sa page sur le site web de Biblissima+.
- Cette page donne accès à l'entrepôt de données et de logiciels du cluster.
- Chaque cluster se réunit une fois par an pendant une semaine de formation et discussions intensives permettant de faire le point, de partager résultats et problèmes, et de fixer une feuille de route commune.
- Les journées Biblissima+ de bilan et prospective, chaque année, permettent aux clusters de confronter leurs résultats, de travailler sur l'articulation des clusters les uns avec les autres, et de déterminer la feuille de route de l'année suivante. Elles ont lieu juste avant le Conseil scientifique, afin que ses membres puissent s'y joindre.
- Le site Biblissima+ traduit par sa présentation et son ergonomie les évolutions du chaînage des outils.

L'intégralité des productions des clusters Biblissima+ est en open access et open source.

❦

B-I.1/ Acquisition des corpus de sources interopérables

Implication : équipe Biblissima+, GED, IRHT, AOROC, CESCM, SAPRAT, SIAF, avec Huma-Num et le Campus Condorcet

B-I.1.1/ La numérisation des sources

Implication : IRHT, GED, équipe Biblissima+, en lien avec POPS

Les partenaires ont déjà numérisé les sources primaires et les publications par millions. Cependant, ce travail de numérisation doit être régulièrement amplifié, corpus par corpus. L'objectif de l'infrastructure n'est pas de financer la numérisation, si ce n'est pour des corpus ciblés dans le cadre des aap partenariaux, tout spécialement dans leur dimension internationale et dans le cadre des environnements de travail proposés par les partenaires. En revanche, favoriser ce genre d'opération fait partie de ses objectifs, grâce d'abord à du matériel. Biblissima+ associe un opérateur majeur de la numérisation des manuscrits, **l'IRHT, et le GED**, qui aura également un rôle très important à jouer dans la numérisation des archives de la recherche en SHS et de collections imprimées pour les usagers. Ces deux opérateurs **souhaitent travailler ensemble à la faveur des synergies créées par le Campus Condorcet, mais aussi en vue de la réalisation des projets partenariaux** Biblissima+. Il faudrait, dans la perspective en particulier de l'association de bibliothèques étrangères aux projets partenariaux mais aussi de n'importe quelle institution de conservation sur le territoire français, acquérir une **station de numérisation légère, mobile**, de façon à pouvoir intervenir dans des bibliothèques, archives ou musées, y compris les plus petits dépôts et pour les plus petits objets, pour des opérations courtes, concernant peu de documents, en particulier dans le cas de documents très dispersés (le modèle est celui, très réussi, de la bibliothèque virtuelle de l'abbaye de Lorsch, qui a numérisé de cette façon quelque 300 manuscrits conservés dans 73 bibliothèques et archives). Ce matériel pourra être utilisé par tous les partenaires, et pourra être hébergé dans les espaces de prise de vue du GED. L'offre du Campus en matière de numérisation sera également renforcée : outre le marché de numérisation du GED, il faudrait, dans la perspective de renforcer les actions de numérisation conjointes, se doter de matériels spécifiques permettant de répondre notamment aux problématiques de **numérisation des archives scientifiques patrimoniales**. En lien avec le projet Equipex+ déposé par Persée et soutenu par le réseau des bibliothèques universitaires et de recherche, l'outil NumaHop pourra être utilisé pour faciliter le processus de numérisation, en permettant un gain de temps et de visibilité à chaque étape.

Comme évoqué plus haut, **l'extension du périmètre de Biblissima+ aux services d'archives**, et en particulier aux services d'Archives départementales, offre l'opportunité de mettre en valeur leurs riches fonds médiévaux, les inventaires révolutionnaires et les inventaires après décès et d'en favoriser la description et la numérisation dans des formats directement exploitables par le portail. Ces services disposent déjà d'une forte expérience en matière de numérisation et de traitement de données, qui s'appuie sur des outils (stations de numérisation) et des compétences en interne ou sur des marchés de numérisation avec des prestataires de qualité. L'apport de Biblissima+ pourrait se situer en amont de la phase de numérisation, afin d'accélérer la préparation de ces opérations (préparation matérielle des documents) et de permettre la reprise de la description des fonds médiévaux à numériser. Les inventaires et instruments de recherche décrivant ces documents sont en effet souvent anciens (pour beaucoup rédigés au 19e siècle) et peu adaptables à la logique d'une consultation sous format web. L'accompagnement de leur reprise permettrait à Biblissima+ de disposer d'un corpus de données beaucoup plus riches, dont l'intégration dans le portail serait facilitée et plus fructueuse (notamment dans une perspective d'enrichissement sémantique et de rattachement à des référentiels de lieux et de personnes).

B-I.1.2/ La numérisation 3D

Implication : AOROC, CESCM, IRHT, SAPRAT

La numérisation des reliures, des monuments ou objets porteurs de texte, des sceaux et monnaies, devra **intégrer la 3D** afin de multiplier les possibilités d'examen et de comparaison des objets numérisés. En épigraphie, pour les textes devenus fantômes, illisibles du fait de leur détérioration ou de la superposition de plusieurs couches d'écritures ou d'images, notamment les graffitis, il faut multiplier les techniques de prise d'images, en utilisant le scan 3D expérimenté dans l'étude de l'art préhistorique et l'archéologie du bâti, mais aussi la photomécanique et les propriétés de la lumière laser développées par l'Institut Pprime (P') (CNRS, UPR 3346, en partenariat avec l'ISAE-ENSMA et l'Université de Poitiers). Plusieurs expérimentations ont déjà été faites, notamment en lien avec une thèse en cours au CESCM sur les graffitis dans les lieux saints en Palestine au Moyen Âge, et elles continueront dans le cadre de **l'ERC GRAPH-EAST**, qui a été obtenue en septembre 2020.

Le programme **SIGILLA** (base numérique de sceaux conservés en France), volet du **portail e-Signa** (EPHE-PSL) s'appuie déjà sur la technologie 3D (scan 3D

portable ARTEC) pour numériser les sceaux les plus précieux ou les plus remarquables. Cette technologie permet des restitutions visuelles et impressions 3D et sera d'un grand secours dans le développement prochain de logiciels de reconnaissance automatique des caractères (applicable à beaucoup d'autres supports écrits comme les monnaies), l'analyse des techniques de gravure, la datation, l'étude des supports, etc.

L'expérience d'AOROC sur la reconstitution 3D d'objets (collection de tablettes cunéiformes de l'EPHE, table d'Héraclée, casque d'Agris, armes du musée, corpus des monnaies celtiques) par le biais d'un scanner 3D Aicon à haute précision, de monuments (Villa de Diomède à Pompéi) ou de sites archéologiques (nécropoles tardo-antiques de Saint-Bertrand-de Comminges, sanctuaire hellénistique et romain de Labraunda en Turquie) par des techniques photogrammétriques, pourra être mobilisée par les chercheurs qui auront à reconstituer des **textes transmis dans un contexte monumental** (inscriptions latines de Rome conservées au Louvre, de Lyon, de Mésie Supérieure, replacées in situ), et éventuellement pour d'autres usages (par exemple les graffitis des prisonniers de Loches et de la forteresse de Chinon). La réalisation de copies (souvent réduites) de ces inscriptions gravées ou peintes, par le biais d'une imprimante 3D, permettra d'ouvrir de nouvelles perspectives de recherches, permettant de mieux saisir la stratification philologique des textes gravés ou peints (comme le laisse entendre l'étude récente de la table d'Héraclée). Elle pourrait permettre de résoudre des questions anciennes comme celle des dédicaces énigmatiques de certains monuments insignes comme le temple dit d'Auguste et Livie à Vienne par l'utilisation de drones et un traitement photogrammétrique. L'achat d'un scanner 3D à main de terrain permettrait d'affiner la photogrammétrie des inscriptions tout en facilitant leur géoréférencement.

B-I.1.3/ L'interopérabilité des images : service IIIF 360 de Biblissima+

Implication : équipe Biblissima+, SIAF, avec un recrutement spécifique pour répandre le protocole d'interopérabilité dans le réseau des Archives de France, avec le Campus Condorcet et Huma-Num

Il est crucial que toutes les campagnes de numérisation veillent désormais à l'adoption du **protocole d'interopérabilité des images IIIF (*International Image Interoperability Framework™*)** créé par l'Université de Stanford (https://iiif.biblissima.fr), qui permet d'appeler dynamiquement, dans un visualiseur web, des images de n'importe quelle bibliothèque numérique dans le monde pour les comparer, reconstituer des documents fragmentaires ou dispersés. Ce protocole est utilisé par le portail Biblissima et IIIF Collections, l'équipe Biblissima en étant le principal promoteur en France. Fin 2018, **Biblissima a créé avec le Campus Condorcet et la TGIR Huma-Num, à titre expérimental, une offre d'expertise et un service d'accompagnement technique, d'hébergement et de développement** à destination des établissements d'enseignement supérieur et de recherche, ainsi que des institutions patrimoniales, **IIIF 360**. IIIF 360 a conseillé et assisté en 2019 **31 projets**, provenant majoritairement de l'ESR. Il s'agit :

- d'un accompagnement technique ou méthodologique en vue de l'implémentation des technologies IIIF dans les bibliothèques numériques (existantes ou en projet) ;
- d'une aide dans le choix d'outils compatibles IIIF (visualiseurs, serveurs d'images, outils d'annotations etc.)
- de solutions de diffusion sur le Web des images numériques (ex : machines virtuelles pré-configurées avec un serveur d'images IIIF en 2020)
- d'outils de traitement d'images, notamment les bibliothèques logicielles permettant d'exploiter le format JPEG2000 (sous licence libre : OpenJPEG ; sous licence commerciale : Kakadu) disponibles sur l'infrastructure de la TGIR Huma-Num.

L'un des besoins le plus souvent exprimés concerne **l'alliance de IIIF et d'Omeka / Omeka S** (version web sémantique d'Omeka, outil de développement de bibliothèques numériques). C'est au niveau de Nakala et de Nakalona (couple Nakala/Omeka) d'Huma-Num que se situent les besoins en prestation de développement de IIIF 360. Au niveau de Biblissima+, il s'agira surtout de développer des modules d'annotation et d'ajouter des fonctionnalités IIIF à Omeka S.

C'est aussi dans le cadre de ce service IIIF 360 que Biblissima+ développera son **pack IIIF – Omeka S, ainsi qu'une version bureau du visualiseur Mirador 3**, installable sur un ordinateur personnel.

L'extension de Biblissima+ au secteur des archives permettra également d'accompagner la diffusion de IIIF à une nouvelle catégorie de producteurs d'images au fort potentiel pour la richesse du portail et de ses outils.

Les équipes proposent trois types de prestations :

- niveau 1, l'offre basique : conseil / accompagnement ; prise en charge gracieusement en best-effort par le consortium des trois acteurs, sans participation des bénéficiaires, mais avec une contrepartie attendue : dans la communication des

bénéficiaires, mention de l'aide apportée par le consortium ;

- niveau 2, l'offre enrichie : conseil / accompagnement + hébergement technique générique ; également prise en charge gracieusement en best-effort par le consortium des trois acteurs, sans participation des bénéficiaires, mais avec une contrepartie attendue : dans la communication des bénéficiaires, mention de l'aide apportée par le consortium ; respect des *FAIR principles* (*Findable*, *Accessible*, *Interoperable*, *and Reusable*) et des pré-requis techniques d'hébergement ;
- niveau 3, l'offre complète : conseil / accompagnement + hébergement technique générique + développements spécifiques; donne lieu à un partenariat formalisé avec participation financière des bénéficiaires (ex : convention), ou à un partenariat subventionné avec partage des aides (ex : dans le cadre d'un projet ANR, ERC) ; avec une contrepartie attendue : dans la communication des bénéficiaires, mention de l'aide apportée par le consortium + respect des *FAIR principles*, des pré-requis techniques d'hébergement et engagement de publication et documentation du code des développements spécifiques sous licence libre.

IIIF 360 sera ainsi une plateforme de services partagée entre Biblissima+ et Huma-Num, qui, avec le soutien du Campus Condorcet, créera une chaîne de traitement efficace duplicable pour d'autres projets qui se feront jour au fil de la vie du programme Biblissima+.

B-I.2/ Prise en compte et cherchabilité des données d'analyse des matériaux

Implication : CRC, avec IRHT, CESCM, AOROC, SAPRAT, équipe Biblissima+

B-I.2.1/ Analyse des matériaux : acquisition et archivage des données

Implication : CRC, IRHT, CESCM, AOROC, SAPRAT, équipe Biblissima+

L'analyse des matériaux des manuscrits (support végétal ou animal comme le papier ou le parchemin, encres, matières colorantes) est un domaine aujourd'hui très prometteur, qui est en train de révolutionner notre approche de l'histoire des bibliothèques anciennes. De telles analyses peuvent permettre de mettre des artefacts en relation les uns avec les autres, d'affiner des datations, de recomposer des réseaux de fabrication et de diffusion. Les exemples d'études à l'interface des sciences expérimentales et des sciences humaines et sociales se multiplient comme le projet « patrimoine écrit » mené actuellement par la Bibliothèque municipale d'Avranches, le **CRAHAM** et le **CRC** sur l'analyse des manuscrits du Mont Saint-Michel ou la thèse réalisée à **l'IRHT** grâce à un projet Biblissima sur les **reliures** de l'abbaye cistercienne de Clairvaux. Les questions de matérialité interrogent de plus en plus les historiens. Des institutions comme l'IRHT, l'EPHE-PSL ou encore l'atelier d'épigraphie médiévale de **Poitiers** sont intéressées par l'analyse des pigments et des techniques de fabrication des objets patrimoniaux. Les philologues d'**AOROC** et ses ingénieurs du CNRS ont développé une approche archéométrique des textes, notamment en portant une attention particulière à la recette des encres utilisées (notamment pour l'étude des cartes d'Albi et de celle figurant sur le célèbre "bouclier de Doura Europos" conservé à la BnF) pour déterminer la provenance et la méthode de fabrication des premières cartes. Le recours aux spectomètres portables XRF (comme les Bruker Tracer 5 et Titan 1) permettaient d'apporter sur ce point des éclairages inédits. Il appliquera à l'écrit des techniques d'analyse bien connues des archéomètres d'AOROC. Cela placera la plateforme Biblissima+ à la pointe de la recherche scientifique appliquée à l'analyse des manuscrits et plus largement à l'ensemble des sources anciennes (par exemple épigraphiques et numismatiques).

Actuellement, une communauté regroupant plusieurs laboratoires en France étudie la matérialité des manuscrits grâce aux sciences analytiques. Les données produites sont de natures très variées, ne présentent pas de structuration commune et sont exploitées de façon hétérogène dans des institutions différentes. Souvent, lors de la conception des projets, aucun partage global à l'extérieur de l'institution n'est prévu. Pour ces raisons, il n'existe aujourd'hui aucun catalogue, aucune bibliothèque numérique en ligne, à notre connaissance, qui mette à disposition des historiens des cultures écrites des données d'analyse des matériaux des manuscrits, des inscriptions, des sceaux etc. **Avec Huma-Num**, Biblissima+ voudrait favoriser l'archivage et la réutilisation de ce type de données, car les bibliothèques détentrices de fonds manuscrits sont régulièrement sollicitées pour des analyses de ce genre et parfois sur les mêmes manuscrits précieux à quelques années de distance, ce qui pose des problèmes de conservation. **Il faut donc rendre cherchables les données d'analyse des matériaux, et pour cela créer l'interface avec la communauté des historiens**.

Le CRC, par sa connaissance des matériaux des manuscrits et les données issues de ses différents projets, est un acteur majeur de la communauté des sciences analytiques et peut jouer ce **rôle d'interface avec la communauté des historiens** des cultures écrites afin de s'inscrire dans une démarche de science

ouverte. L'IRHT, comme AOROC, interagira avec le CRC pour que soit développée **une méthodologie d'archivage des données**. L'IRHT pourra fournir dans certains domaines des sets de données : par ex. les résultats concernant plus de 500 manuscrits de la Genizah du Caire, conservés à la bibliothèque de Cambridge (Angleterre). Sur ces problématiques d'archivage, l'échange avec l'équipex+ ESPADON sera de première importance : c'est l'un des domaines d'articulation forte entre les deux projets, qui ont échangé des lettres d'intention.

B-I.2.2/ Analyse des matériaux : « cherchabilité » des données

Implication : CRC, équipe Biblissima+

Il s'agit d'arriver à une **structuration et une normalisation des données** respectant des modèles communs afin de permettre le stockage, le partage et l'interrogation des données de la communauté. Biblissima+ offre une opportunité unique de mener, avec l'aide d'un informaticien, une réflexion sur les outils informatiques nécessaires à leur exploitation. Ces outils se doivent d'être **simples d'utilisation, fiables et accessibles à des non-spécialistes** des matériaux. Le développement de ces outils représente un véritable défi. En effet, la complexité des informations produites peut être très variable et l'exploitation des données plus ou moins aisée pour des non-spécialistes.

Il faudra donc **évaluer le niveau de complexité de l'information fournie** : données brutes, données exploitées et/ou interprétées. La seule mise à disposition des données brutes ne peut suffire à permettre l'accessibilité à tous les niveaux d'informations. Au-delà de la question des formats propriétaires (formats qui peuvent être lus uniquement à l'aide d'un logiciel payant), qui peut souvent se résoudre par une conversion des données, l'exploitation de ces données nécessite souvent des compétences spécifiques qu'un non-spécialiste ne maîtrisera pas ou pas totalement. Cependant, **donner des informations déjà traitées entraîne le risque de donner une information partielle** (qui peut par exemple répondre à la problématique que se posait le chercheur ayant traité les données mais ne répond pas aux nouvelles questions de la personne qui consulte ces données) ou dont l'interprétation forcément subjective peut être remise en question à la lumière de nouvelles données. C'est souvent le cas lorsqu'on se confronte à l'imagerie chimique (imagerie hyperspectrale, scanner de fluorescence des rayons X…), ces données complexes qui produisent des spectres en chaque pixel d'une image et qui peuvent être traitées soit comme de l'imagerie soit comme spectroscopie. Le risque de produire des images induisant des interprétations erronées est grand. Par exemple, grâce à l'imageur de fluorescence de rayons X, on peut obtenir des cartographies des éléments présents dans une enluminure. Une cartographie du cuivre où l'intensité du pixel est corrélée à la concentration en cet élément est facile à obtenir. Il serait alors aisé de penser que cette cartographie présente l'intégralité des données à un lecteur sans que ce dernier ait à générer cette cartographie**. Cependant des biais sont possibles, liés à la dynamique de l'image**. En effet, si deux pigments contenant du cuivre sont présents dans cette enluminure, par exemple un pigment vert au cuivre comme la malachite et un autre comme un pigment bleu outremer présentant le cuivre en très faible quantité (c'est un silicate et le cuivre s'y trouve sous forme d'impureté), il est fort probable que seul le cuivre présent dans le pigment vert apparaisse, le cuivre du pigment bleu n'étant présent qu'en très faible quantité. Pourtant cette information pourrait se révéler très importante car potentiellement reliée à la provenance de ce pigment bleu. Cette information ne pourra être mise en évidence qu'en retraitant les données brutes et en modifiant les paramètres permettant de générer la cartographie (par exemple en ne tenant pas compte des pixels dont la quantité de cuivre est trop importante ou en modifiant l'échelle de représentation des concentrations). Une analyse superficielle des cartographies déjà générées aurait malheureusement conclu à l'absence de cuivre dans ce pigment bleu.

Le cas de l'imagerie hyperspectrale est également complexe car les outils mathématiques permettant de générer les cartographies de certains matériaux ne sont pas tous adaptés et **il est parfois difficile de mettre en évidence par une seule image ce qu'il est possible de voir sur les spectres**. L'hydrocérusite (une forme de blanc de plomb) se traduit par exemple par un très faible pic d'absorption visible sur les spectres de réflectance mais il est très difficile à faire ressortir de façon certaine sur une cartographie… Enfin **une mauvaise interprétation est toujours possible**. Par exemple, on peut interpréter la présence de plomb dans un pigment orangé comme la preuve de la présence de minium (un oxyde de plomb), mais en l'absence de technique confirmant cette interprétation (par exemple la spectroscopie Raman qui donne une information ici plus spécifique), il est également possible que le signal du plomb provienne de la présence de blanc de plomb et que la couleur orange provienne d'un autre pigment ou colorant.

Ainsi, la fiabilité de l'information est un point essentiel mais qui se heurte à la complexité des matériaux du patrimoine. Un processus progressif de mise à disposition des données pourra être envisagé avec **une première étape incluant des jeux de données simples d'exploitation** (par exemple les données issues de la photographie scientifique ou ne faisant intervenir qu'un type de technique sur tout un

corpus de documents) afin de **tester la faisabilité du concept avant de se confronter à des données plus complexes qui amèneront au développement d'outils spécifiques**.

Durant la première phase de Biblissima+, le CRC pourra **utiliser des jeux de données déjà existants pour tester la pertinence** de ces outils, et évaluer le travail de préparation des données (tri, renseignement des métadonnées…), afin de les rendre accessibles à tous, réutilisables et comparables à de nouvelles acquisitions. Ce travail de préparation des données est primordial pour qu'une personne n'ayant pas acquis les données puisse les réutiliser. Par exemple, **toute analyse ponctuelle (réalisée en un point précis de l'objet) doit être localisée précisément et les paramètres d'acquisition (notamment la taille de la surface analysée) renseignés**. L'expérience du CRC sur les manuscrits du Mont Saint-Michel montre par exemple une certaine hétérogénéité des matériaux utilisés pour des couleurs apparaissant aujourd'hui de façon très similaire (vermillon ou minium pour des teintes rouge-orangées très semblables ou des verts de compositions différentes avec ou sans zinc au sein de la même enluminure). Il n'est donc pas si évident de conclure sur la nature de tous les matériaux d'un même document à partir de quelques points d'analyse. **Le tri des données** est également important car une analyse dont la fiabilité n'est pas certaine ne devra pas être transmise (ou alors en ajoutant une information sur les doutes potentiels). Afin de se placer dans une démarche pérenne**, l'équilibre entre le volume d'informations à mettre à disposition et l'intérêt de l'information transmise devra être trouvé**.

Cependant si ce travail de préparation est trop lourd, on peut penser que la procédure mise en place dans ce projet ne sera pas ou peu appliquée par les chercheurs ayant acquis un grand nombre de données. Il est donc important de réfléchir à une stratégie qui reste réalisable et ne soit pas trop chronophage. Le temps nécessaire à cette étape devra par la suite être intégré dans les projets à venir…

Les nouveaux projets de recherche ciblés sur les collections de manuscrits qui seront menés au sein du CRC durant cette première phase seront autant d'opportunités de compléter ce travail, en réfléchissant à la meilleure façon d'organiser les données, non pas seulement dans une démarche de restitution a posteriori, mais aussi au moment de leur collecte. **L'objectif de la première phase de Biblissima+ est de construire des outils qui pourront être utilisés dans la seconde phase par tout laboratoire produisant ce type de données et souhaitant alimenter les bases de données communes**. Des accords pourraient être passés avec des détenteurs de bases de données sur les matériaux du livre ancien, par ex. l'équipe du projet Beast2Craft en Angleterre sur l'identification d'espèces animales par l'analyse des protéines et de l'ADN, ou les bases de dendrochronologie. En France, un partenariat avec l'IRAMAT pourrait également être envisagé.

B-I.3/ Intelligence artificielle, reconnaissance de formes et d'écritures manuscrites

Implication : IRHT, TEKLIA, CJM, eScripta et PSL, AN, AOROC, CRAHAM, SAPRAT, CESR

L'analyse automatisée de formes et donc de mise en page et d'écriture à l'aide de l'intelligence artificielle est l'un des domaines à la pointe de l'innovation actuellement. Quatre des partenaires de Biblissima+ sont particulièrement investis : PSL, en particulier l'EPHE-PSL et l'ENC-PSL, avec le programme e-Scripta et l'outil eScriptorium, ainsi que l'IRHT, en partenariat avec les AN. Ces partenaires ont développé, pour l'instant largement séparément, des technologies de transcription automatique grâce à des outils HTR comme Arkindex (IRHT et Teklia) et Kraken (en partenariat avec l'EPHE-PSL), et en établissant des ponts avec Transkribus (les partenaires de l'IRHT dans les projets HIMANIS et HOME sont membres du consortium READ qui a développé Transkribus).

Il faudra **encourager les rapprochements entre les développements complémentaires** de Kraken et Arkindex. Kraken, entièrement open source (licence Apache 2) donne aux chercheurs la possibilité d'entraîner ou réentraîner la machine selon leurs besoins et problématiques, par ex. pour les écritures rares et variées, mais cela implique plus de travail et d'expertise au début, alors qu'Arkindex est déjà entraîné et donc plus facile à utiliser, mais moins adaptable aux besoins différents. Arkindex est une architecture prévue pour des traitements de masse, le suivi de production (indicateurs de qualité, supervision) et le service aux utilisateurs, avec des fonctionnalités linguistiques. Kraken est optimisé pour l'analyse fine d'écritures variées et la réalisation des tâches scientifiques par les chercheurs individuels ou en petite équipe.

L'automatisation de l'analyse de la mise en page et de la transcription dans tous les systèmes graphiques a de multiples implications, et elle intéresse autant les spécialistes du manuscrit que les spécialistes de l'imprimé ou de l'épigraphie, et, au-delà, la société tout entière. Les outils d'HTR développés pour toutes les écritures pourraient répondre aux besoins de diverses communautés fréquentant le portail Biblissima+ ou que l'on voudrait amener vers le portail :

– **pour les conservateurs**, le pré-catalogage automatisé des collections d'artefacts numérisés, particulièrement important dans les situations d'urgence, par exemple de sauvetage d'une collection ancienne dans un pays en guerre, mais pas seulement.

Il serait intéressant à cet égard, par ex., de développer des algorithmes de calcul automatique de la taille des manuscrits et de leur mise en page, et de générer automatiquement des tables de matières.

– **pour les chercheurs avancés**, la transcription rapide de grands corpus, en particulier à des fins de comparaison et parfois de restauration des textes (cas des textes effacés, fragmentaires, endommagés) ; la reconnaissance d'entités nommées (cote, noms de personnes, titres d'œuvres) et l'alignement sur des référentiels : dans les métadonnées (en vue de la liaison entre les bases) et dans le plein texte des manuscrits numérisés, comme cela se fait dans le projet HOME sur les cartulaires ;

– **pour des internautes novices**, qui abordent par exemple les écritures manuscrites mais ne savent pas encore bien les déchiffrer, ou qui font des recherches de généalogie, un affichage à la volée de la transcription du manuscrit qu'ils consultent;

– **pour tous,** l'un des grands souhaits irréalisables à ce jour est de pouvoir chercher un mot, un nom, en plein texte dans un document ancien quel qu'il soit, ce qui suppose un passage de HTR avancé sur le manuscrit puis un lien entre cette transcription et les zones pertinentes de l'image, afin que l'internaute soit guidé vers le bon feuillet et la bonne zone de l'image pour trouver le texte qui l'intéresse.

– Ces avancées technologiques pourraient favoriser, avec une bonne conception de l'espace d'annotation décrit plus bas, des entreprises de **science participative** : plusieurs partenaires évoquent cette piste prometteuse, par exemple pour le relevé **d'inscriptions latines** modernes sur les monuments (CJM), ou pour les **inventaires anciens de manuscrits et d'archives** (IRHT), qui pourraient intéresser les associations d'histoire locale – sans compter la possibilité, en mobilisant ce lien entre science et société, de retrouver des documents en main privée. On pourrait utiliser, pour guider les participants, une **cartographie interactive des sources encore inexploitées.**

D'autres partenaires utilisent ces outils ou peuvent les utiliser en leur soumettant de nouveaux défis. Le programme BVH du CESR souhaite faciliter l'accès à d'autres corpus du français préclassique et **répondre aux besoins des bibliothèques, des archivistes et des chercheurs pour la lecture en texte intégral**, en mettant en œuvre des compétences linguistiques en amont et dans la phase de correction (**outil Pyrrha du CJM**). Dans cette phase, on utilise les dictionnaires d'entités nommées déjà balisées en TEI dans un corpus annoté afin de les repérer automatiquement dans de nouveaux textes. L'équipe du *Corpus des inscriptions de la France médiévale* à Poitiers est intéressée également, car il y a **un enjeu pour l'épigraphie** : ce problème est actuellement traité par Kraken et eScriptorium, qui sont appliqués aux inscriptions en vieux vietnamien dans le cadre du projet **Vietnamica** (EPHE-PSL) soutenu par une « Advanced Grant » de l'ERC. Autre exemple d'usage : Le CJM (ENC-PSL) et les AN sont partie prenante d'un projet ANR déposé par le Lamop (CNRS - Paris 1) sur les **registres capitulaires de Notre-Dame de Paris**. Le projet comprend de l'acquisition automatique de texte par HTR, de la reconnaissance d'entités nommées, de l'édition en XML TEI. Autre exemple: LectauRep, un projet de l'INRIA (ALMAnaCH) avec l'EPHE et les Archives Nationales, concerne la transcription automatique des **répertoires des études notariales** (le minutier central) des 19e et 20e siècles. Il utilise Kraken et eScriptorium sur un corpus d'environ 2000 registres (300 à 500 pages chacun). Un partenariat entre l'Université de Tours (CESR et LIFAT), ENC-PSL, EPHE-PSL, INRIA (ALMAnaCH), Université de Neuchâtel, se met en place pour le développement des projets scientifiques centrés sur l'entraînement de la machine (machine learning) pour **l'automatisation de la chaîne éditoriale des livres anciens**.

B-I.3.1/ Les recherches de l'IRHT et TEKLIA sur les textes latins et français

Implication : IRHT, TEKLIA

A l'IRHT, Dominique Stutzmann a développé avec la société TEKLIA la plate-forme **Arkindex**, utilisée par plusieurs projets de recherche sur les livres médiévaux, basée sur l'intelligence artificielle et la vision par ordinateur pour alimenter l'analyse historique et visuelle.

Les principales fonctionnalités sont les suivantes :

- import des ressources numérisées interopérables (IIIF) ;
- classification automatisée des pages (reliure, page blanche, miniature, miniature et texte, calendrier, etc. ;
- segmentation automatisée de la mise en page (page, colonne, lignes) ;
- classification automatisée des éléments visuels (miniature, initiale historiée, initiale décorée, remplissage de ligne, notation musicale, etc. ;
- reconnaissance automatique des textes manuscrits (couvrant le moyen français et le latin) ;
- reconnaissance automatique des entités nommées ;
- identification automatisée du texte ;
- navigation en fonction de la structure intellectuelle (texte ou parties de texte) ou de la structure physique (page) ;
- interface d'annotation.

La plate-forme est actuellement utilisée pour deux projets : HORAE – « Heures : Reconnaissance, Analyse, Edition » (ANR-17-CE38-0008), dans lequel des livres d'heures subissent une analyse visuelle (nombre et taille des miniatures, nombre de lignes etc.) et sont "lus" par le système Arkindex et leur texte analysé pour produire une liste d'unités textuelles minimales

telles que psaumes, antiennes, leçons, etc. et une table des matières détaillée (par exemple Heures de la Vierge > Matines > 1er Nocturne > 1ère Leçon…) – HOME – « History of Medieval Europe » (JPI Cultural Heritage), dans lequel le contenu textuel des cartulaires médiévaux est lu, indexé et analysé pour en extraire toutes les entités nommées (noms de personnes et de lieux, dates). Comparé à d'autres outils disponibles, tels que *Transkribus* pour la reconnaissance de textes manuscrits et la transcription de textes d'une part, ou *Recogito* pour l'extraction d'entités nommées d'autre part, Arkindex offre un flux de travail facile pour l'analyse visuelle et la reconnaissance de textes à grande échelle. La collaboration avec Teklia fournit également une solution robuste avec l'infrastructure et le soutien nécessaires pour un fonctionnement fiable à grande échelle.

B-I.3.2/ Kraken, un module HTR pour toutes les écritures (e-Scripta, EPHE-PSL)

Implication : eScripta en lien avec divers projets internationaux

Kraken est **le premier élément de la suite d'outils conçue par les chercheurs de l'EPHE-PSL au sein de Scripta-PSL pour étudier des corpus de manuscrits écrits dans n'importe quelle langue et n'importe quel système graphique**. Ce formidable défi a été relevé avec la conception de ce moteur autonome, open-source, conçu spécifiquement de manière modulaire et avec le moins de présuppositions possibles sur le système d'écriture. Basé sur les principales méthodes actuelles d'apprentissage approfondi à l'aide de l'intelligence artificielle, il donne des résultats comparables aux meilleurs autres systèmes, mais il est en même temps **plus flexible et personnalisable que les autres logiciels**, il est donc particulièrement bien adapté à la grande variété de types d'écriture que l'on trouve dans les langues "rares" et historiques : qu'elle soit cursive ou non, qu'elle soit écrite sur une ligne de base, à partir d'une ligne supérieure ou le long d'une ligne verticale ; qu'elle aille de droite à gauche, de gauche à droite, de haut en bas, qu'il s'agisse de boustrophedon (écriture qui va de gauche à droite puis de droite à gauche comme le sillon tracé par une charrue), ou un mélange de tous ces éléments ; qu'elle soit sur des lignes courbes, droites, diagonales ou une combinaison de ces éléments ; et ainsi de suite. Un autre élément essentiel ici est que, pour fonctionner avec des écritures et des langues rares, le système doit être capable de s'entraîner avec des quantités relativement faibles de données. Les approches développées par des entreprises telles que Google dépendent de quantités de données extrêmement importantes et ne fonctionnent donc pas bien avec des écritures rares et anciennes.

L'outil est souple : les utilisateurs ont toute liberté pour **définir leurs propres principes de transcription, jusqu'à la translittération complète.** Par exemple, les objets écrits en écriture sémitique peuvent être transcrits en alphabet latin (comme c'est généralement le cas en épigraphie sémitique) ou en arabe ou en hébreu. Les documents turcs ottomans peuvent être transcrits en alphabet latin comme le turc moderne, ou dans le type d'écriture arabe utilisé dans l'Empire ottoman.

Kraken comprend une **archive ouverte de modèles qui fonctionne via Zenodo**. Les utilisateurs peuvent publier leurs modèles, et/ou les télécharger, ce qui leur donne le contrôle de leurs données et des modèles formés. Cela signifie également que les utilisateurs n'ont pas besoin de perdre du temps, des efforts et des ressources informatiques pour réapprendre des modèles qui existent déjà, mais qu'ils peuvent au contraire s'appuyer sur le travail déjà effectué, par exemple en prenant un modèle déjà entraîné pour une écriture et en le réentraînant pour une écriture différente mais similaire.

L'analyse automatique d'images de documents écrits, sur la base de réseaux de neurones convolutifs, permet d'isoler des objets de leur arrière-fond et d'établir une distinction entre l'écriture principale, la décoration (enluminures, lettrines, etc.) et les annotations interlinéaires ou marginales quel que soit le sens de l'écriture. Il est également possible pour les chercheurs de réentraîner Kraken en fonction de leurs besoins particuliers, par exemple pour détecter automatiquement des lignes de texte même avec des mises en page très complexes (comme on en trouve souvent en arabe, par exemple), ou pour réentraîner le modèle d'HTR pour un corpus, un style d'écriture ou un scribe particulier. Kraken est déjà adapté pour les manuscrits mis à disposition par IIIF via eScriptorium (voir ci-dessous), mais il faudrait développer le lien entre la constitution de corpus dans Biblissima+ et leur traitement par les outils d'HTR mis au point par les partenaires. **Biblissima+ s'ouvrant potentiellement à toutes les langues et systèmes graphiques, le programme devrait financer les futurs développements de l'outil Kraken**, et donc assurer la stabilité de l'emploi des ingénieurs qui le développent, actuellement co-financés par l'Université PSL et par le projet infradev Resilience (dir. Denis Pelletier pour la France). Il faudra aussi assurer le renouvellement, dans 5 ans, de l'infrastructure de calcul financée par le DIM STCN de la région Île-de-France avec un cofinancement provisoire de l'EPHE-PSL, de l'IRHT, et du projet *Sofer Mahir* de Daniel Stökl Ben Ezra, et hébergée par MesoPSL à l'Observatoire de Paris - PSL (8 gpu et 1 Po de stockage).

B-I.3.3/ Reconnaissance des filigranes, d'éléments du décor, héraldiques et numismatiques

Implication : IRHT et le CJM avec l'INRIA et l'Ecole des Ponts, PSL, AN, AOROC, CRAHAM, SAPRAT, CESR

Les filigranes permettent de dater et localiser les papiers anciens, et donc, dans une certaine mesure, les écrits qu'ils portent. Il en existe des répertoires imprimés (Briquet est le plus commun), mais d'usage peu commode. Le relevé des filigranes dans les livres anciens n'est pas aisé non plus. Il y a donc là un enjeu important pour les historiens du livre et des textes. **L'ENC-PSL, l'IRHT, l'INRIA et l'École des Ponts** ont développé dans le cadre de **l'IRIS de PSL *Sciences des données, Données de la science*** le projet « Filigranes pour tous », qui a débouché sur une application web et mobile permettant d'accéder au répertoire de Briquet et, sur la base d'un corpus test des **AN (Minutier central)**, d'identifier les filigranes photographiés par un smartphone. **L'application est conçue pour être participative**, c'est-à-dire enrichie par tout un chacun lors des consultations en archives.

Il faut maintenant aller plus loin que la preuve de concept, en **améliorant l'innovation, en la diffusant, en élargissant** enfin son champ d'application. Il est désormais nécessaire d'alimenter la base en métadonnées. En l'état, elle peut prendre en compte les filigranes des manuscrits sur papier des bibliothèques de France, très mal renseignés mais qui feront dorénavant l'objet de campagnes de reproduction par caméra infrarouge. **L'enrichissement viendra aussi du crowdsourcing, ce qui implique des opérations de validation scientifique**. Au-delà, le corpus augmenté et les répertoires anciens peuvent être mis en interopérabilité avec d'autres initiatives internationales, par exemple celle du réseau WZIS (Wasserzeichen-Informationssystem, https://www.wasserzeichen-online.de) auquel **l'IRHT** est affilié et qui a récemment intégré des filigranes des manuscrits de Saint-Omer. Des journées de formation peuvent aussi être organisées, sur la collation matérielle et l'exploitation informatisée des filigranes.

Autres domaines dans lesquels la reconnaissance de formes est en plein essor, et prendra rapidement en compte l'intelligence artificielle : **la numismatique et la sigillographie**.

Le **CRAHAM** se propose d'enrichir la reconnaissance d'éléments numismatiques, c'est-à-dire participer à l'effort de **référencement et d'indexation d'empreintes représentant des monnaies ou des types monétaires**. L'inspiration monétaire sur des méreaux de plomb ou des jetons est attendue et surtout documentée. On insistera davantage sur la reprise, attestée depuis l'Antiquité, de types monétaires sur des sources archéologiques comme la céramique en particulier (vases, lampes, tuiles, amphores, hosties, etc.). Cette documentation n'a jamais été renseignée de manière systématique et pourrait faire l'objet de campagnes de reproductions – qui seraient complémentaires à un inventaire des types monétaires représentés sur les sceaux. La collation et l'exploitation de ces sources permettraient d'en savoir plus sur les représentations choisies au fil des siècles. De son côté, **AOROC** a entrepris depuis 5 ans l'établissement d'un **corpus 3D de monnaies** celtiques et italiques (pré-romaines) à l'aide d'un scanner 3D de grande précision (Eikon, acquis sur des crédits PSL à l'ENS en 2017), en collaboration avec l'Ecole française de Rome et le musée archéologique de Brescia, en portant une attention particulière aux décors figurés et aux coins, dont la combinaison est l'expression d'un langage monétaire de sociétés sans écriture. Cette langue non verbale est identifiée par la mise en série de très nombreuses pièces. Sa régionalisation et sa chronologie sont aussi mises en évidence par leur insertion dans une base de données (base Fer) et leur géoréférencement sur un Web SIG conçu et développé par AOROC, en collaboration avec la société Geocarta, ancienne start-up du CNRS, lancée par le laboratoire. Biblissima+ permettrait de poursuivre l'acquisition des données et leur mise en ligne sur des outils assurant le libre accès (Chronocarto, sketchfab, Nakala/HumaNum).

La reconnaissance des formes trouve un terrain d'application privilégié dans la **reconnaissance d'armoiries et de signes héraldiques** en général. Dans e-Signa (EPHE-PSL), les données emblématiques sont issues des bases associées où elles sont collectées via un module héraldique commun aux trois bases. Ce module, déjà existant, permet, en back office, de formaliser les informations héraldiques et de les restituer sous forme d'un dessin normalisé et d'une description standardisée. En front office, l'**interface de recherches héraldiques** permettra d'interroger par image ou texte les ressources du portail à l'aide d'un **outil simple de reconstruction graphique**. Les procédés de reconnaissance automatique d'image, appliqués aux armoiries, connaissent actuellement un essor important porté par le Pr. Torsten Hiltmann de l'université Humboldt de Berlin (ancien étudiant de la Training School IRHT-BIBLISSIMA-SISMEL de 2014). Cet outil pourra être agrégé à l'interface héraldique d'e-Signa dans un futur proche.

Une méthodologie très proche est en cours de développement à **PSL, à AOROC**, pour les monnaies, en collaboration avec **Mines ParisTech**. Cette recherche donne lieu depuis 2018 à une thèse interdisciplinaire financée par PSL (Sofiane Hourache, sous la direction de Thierry Lejars DR2 CNRS d'AOROC et François Goulette Professeur à Mines ParisTech). L'outil qui donnera lieu à la création d'un logiciel libre d'accès viendra enrichir les capacités d'analyse de Biblissima+.

En marge du projet Filigranes pour tous a été développée une application, **Extractor**, **permettant d'isoler dans un manuscrit donné tous les éléments de décor**, puis de comparer ces éléments d'un manuscrit à l'autre. Un tel outil est particulièrement précieux pour l'étude, par exemple, des lettres filigranées. **Il doit pouvoir donner lieu à un outil exploitable par les bibliothèques numériques**, sous forme d'un plug-in mobilisable à la demande. Les **BVH** participent également à l'effort de référencement et indexation des décors imprimés via la base de données BaTyr (vignette, bandeaux, lettrines), ainsi que des gravures représentant des monnaies, des armoiries, des épigraphes présentées dans le corps des textes ou dans des parties de reliures.

B-I.4/ Traitement approfondi des systèmes graphiques et analyse des documents

Implication : eScripta, AOROC, SAPRAT, PSL

B-I.4.1/ Analyse des écritures anciennes et environnement d'annotation : Archetype et eScriptorium

Implication : eScripta, AOROC

L'analyse des écritures anciennes (l'analyse paléographique) a besoin d'une infrastructure commune, ou du moins d'un moyen **d'aligner les différentes terminologies et méthodes descriptives permettant l'interopérabilité, l'échange et l'interconnexion**. Les chercheurs sont de plus en plus conscients de la valeur et de l'intérêt de la paléographie « transversale », au sens du Groupe de recherches transversales en paléographie (GRTP) qui étudie « des questions de fond de paléographie et d'histoire de l'écrit sans cloisonnement géographique ni chronologique ». Cependant, cette transversalité nécessite une base commune de communication et cela suggère encore la nécessité d'aligner des concepts communs. Il est donc nécessaire de formaliser et d'aligner autant que possible les descriptions des experts qui, si elles réussissent, permettront de telles comparaisons transversales. Il faut pour cela définir une, voire plusieurs ontologies d'écritures et les intégrer dans **un système d'annotation et d'analyse structuré**, ce qui permettrait aux chercheurs d'utiliser des modèles similaires mais différents en fonction de leurs besoins particuliers, mais aussi de rechercher parmi ces différentes instances pour fournir des résultats unifiés. Les ensembles de données produites peuvent ensuite être disponibles en ligne pour les chercheurs, mais peuvent également être disponibles pour une récolte et une récupération automatique par le biais d'une interface de programmation d'application (API). Le logiciel lui-même (sans données), libre, sera également mis à la disposition d'autres utilisateurs pour une installation sur leurs propres serveurs web, mais des versions seront également développées et regroupées dans des **« machines virtuelles »** que chacun pourra télécharger et exécuter à partir d'un ordinateur.

Ce logiciel existe déjà : ***Archetype*, un logiciel libre, gratuit et « Open Source ».** Anciennement connu comme le « DigiPal Framework », il a été conçu dans le but d'offrir une méthodologie visible et reproductible permettant d'explorer les données paléographiques et de les communiquer comme preuves et arguments, ainsi que de gérer et afficher des textes électroniques (telles que des éditions et traductions, des descriptions paléographiques et codicologiques, etc). **Il est déjà utilisé pour une vingtaine de projets sur les écritures manuscrites et épigraphiques en latin, grec, hébreu, vieil anglais et vieux khmer, ainsi que des expériences sur l'arabe, le chinois, le japonais, le cunéiforme, les hiéroglyphes égyptiens et maya, et les brouillons de Marcel Proust**. Archetype a été utilisé également pour l'analyse de la décoration dans les manuscrits hébreux, la peinture de la Renaissance, et la tapisserie de Bayeux. En 2016, il a reçu le prix de la Medieval Academy of America pour le meilleur projet en humanités numériques. Archetype a été téléchargé de DockerHub plus de 1200 fois, en plus des téléchargements de 'code source' de GitHub. Cependant, Archetype a été développé par Peter Stokes (directeur d'études EPHE, PSL) et les équipes informatiques de King's College à Londres. Il dépend d'outils et logiciels développés par des équipes qui ne travaillent plus ensemble : du coup, ils devraient être entièrement redéveloppés pour assurer la pérennité de l'outil au sein d'e-Scripta.

Ce redéveloppement serait l'occasion de l'inclure dans un environnement de travail ergonomique dont la **première brique sera Kraken**, qui n'a pas encore d'interface utilisateur et n'est pas ergonomique, **et la seconde Archetype**. En plus des annotations structurées sur les images pour l'analyse paléographique, le texte lui-même nécessite aussi des **annotations**, généralement sous forme de balisage XML. Il s'agit au minimum d'indiquer des éléments structurels (pages, lignes, paragraphes, chapitres, etc.) mais aussi des détails philologiques tels que des interventions éditoriales, des annotations linguistiques, etc. De nombreux outils existent déjà pour cela, mais ils sont souvent conçus en fonction de systèmes d'écriture spécifiques et ne sont pas facilement adaptables à la grande variété d'écritures rares et anciennes. Ils doivent donc également être intégrés aux autres composants énumérés ici, et permettre l'exportation vers d'autres systèmes en fonction des besoins. Une fois le texte obtenu, les images et le texte annotés et analysés, la tâche finale consiste souvent à publier les textes en ligne. Une fois de plus, l'infrastructure devra **tenir compte des**

nombreuses conventions différentes des différents domaines, ainsi que des **défis posés par les écritures rares et anciennes,** également dans les imprimés annotés de façon anonyme. Là encore, les utilisateurs pourront définir différents formats pour différents balisages, créer différentes visualisations et interactions, etc. En pratique, cela prendra la forme de quelques dispositions de base qui peuvent être personnalisées relativement facilement grâce à une interface graphique visuelle, de quelques personnalisations plus avancées qui nécessiteront un niveau relativement faible de connaissances en programmation, et, pour ceux qui disposent des ressources et des compétences nécessaires, la capacité de définir des visualisations entières grâce à la programmation de feuilles de style XSLT et d'autres plugins.

On a donc besoin d'un logiciel d'infrastructure avec une interface ergonomique pour lier les fonctionnalités de Kraken et Archetype, ainsi que d'autres outils, en particulier pour annoter les textes et publier les résultats, ce qui permettra une intégration plus étroite des résultats HTR et des annotations paléographiques ou autres, ainsi que l'export en format IIIF, TEI et autres formats standards. **eScriptorium fournira donc une interface utilisateur graphique** pour les tâches suivantes (entre autres) :

- Import automatique d'images à partir d'un manifeste IIIF
- Saisie de données de « vérité de base » (« ground truth ») pour entraîner Kraken et pour corriger ses sorties (tracer des lignes de texte sur l'image pour détecter les lignes, ou taper des transcriptions)
- Annotation des images selon la (ou les) ontologie(s) de l'écriture élaborée(s) à partir de celle d'Archetype
- Annotation des textes avec un balisage de TEI XML
- Publication des textes, images et annotations
- Import et export de données dans différents formats standards, tels que les annotations IIIF, ou PAGE, Alto ou TEI XML

Ce système dédié est essentiel pour soutenir l'apprentissage automatique pour un nombre important d'utilisateurs. Avec l'intégration des modèles et ontologies IIIF et Linked Open Data, et l'expertise en apprentissage machine, il permettra pour la première fois l'apprentissage automatique à grande échelle appliqué à des corpus dispersés de manuscrits et de documents, en principe dans n'importe quelle langue ou écriture.

B-I.4.2/ Multipal pour dater, localiser, lire toutes les écritures

Implication : SAPRAT, EPHE-PSL, PSL

Avec le soutien de Scripta-PSL, de la direction de la recherche de PSL et du Groupe de Recherche Transversale en Paléographie (GRTP), des spécialistes de l'EPHE-PSL, du CJM (ENC-PSL) et du Collège de France ont élaboré dès 2018 un tutoriel interactif en paléographie accessible en ligne, intitulé MultiPal. Multidirectionnel, MultiPal sert à apprendre à déchiffrer les manuscrits, documents et inscriptions originaux dans **toute la gamme des écritures et langues de l'Antiquité et du Moyen-Âge**. Cet outil permet de visualiser des images de haute qualité et de transcrire le texte manuscrit mot par mot ou idéogramme par idéogramme, selon les langues, dans des cases prédéfinies. La transcription est vérifiée et corrigée automatiquement. Les utilisateurs ont aussi l'option de visualiser la transcription complète du document. Ils peuvent également consulter une notice descriptive de chaque élément. En libre accès en ligne, le tutoriel permet aux étudiants et à toute personne intéressée de s'initier à la lecture de manuscrits. Pour l'instant, un nombre limité d'exercices de lecture en latin, grec, égyptien, copte, arabe, hébreu, araméen, syriaque, chinois, sanscrit et cyrillique a été mis en place.

La suite du projet envisage deux directions. Le projet ambitionne d'augmenter le nombre d'exercices dans des écritures déjà représentées pour mieux refléter leur étendue géographique et chronologique, et **d'inclure des exercices dans d'autres systèmes d'écriture**. Par ailleurs, il faut faire évoluer le tutoriel de transcription paléographique vers une analyse plus poussée des écritures en ligne, **un Album paléographique** qui servira de base à la **typologie** des écritures et servira de référence accessible aidant à la **datation** des manuscrits. **L'Album sera développé en utilisant le support d'Archetype afin de faciliter l'interaction entre cet outil d'apprentissage et de recherche et l'environnement d'analyse des écritures et de leurs supports**. Multipal sera conçu comme un exemplier exhaustif des échantillons manuscrits datés et datables, représentés par des images de haute qualité, suivi d'une analyse paléographique sur deux niveaux : la vision globale de l'écriture et l'analyse des lettres et leurs composantes du point de vue de leur morphologie et de leur ductus.

B.I.5/ Edition de sources en TEI

PDN-MRSH et CRAHAM, CJM, AN, IRHT, HiSoMA et CIHAM, AOROC, CESCM, CRH, SAPRAT, CESR

L'édition électronique en TEI n'est plus une nouveauté, mais ce qui est exceptionnel dans le

bouquet Biblissima+, c'est la diversité de ses applications et des environnements de travail dans lesquels la TEI joue un rôle déterminant, et le souhait de convergence des partenaires.

B-I.5.1/ TEI et épigraphie, de l'Antiquité à l'époque moderne

Implication : HiSoMA, AOROC, CESCM, CRH, SAPRAT, CESR, CJM

Avec la présence d'HiSoMA , de l'EHESS, d'AOROC, du CESCM, du CESR, du CJM dans l'infrastructure Biblissima+, nous avons le moyen de **développer un pôle d'éditions épigraphiques en TEI exceptionnel, couvrant la longue durée** : de l'Antiquité grecque (HiSoMA , EHESS avec le projet d' « Anthologie du livre grec manuscrit » ALGreM) et latine (AOROC), jusqu'aux antiquités moins traditionnelles utilisant des langues rares (notamment italiques, égyptiennes mais aussi orientales), au projet de science participative sur l'épigraphie latine moderne porté par l'ENC, en passant par le *Corpus des inscriptions de la France médiévale* du CESCM, ou encore les recherches du CESR sur l'épigraphie de la Renaissance et sur les sentences peintes (latines et grecques) du château de Montaigne qui ne sont pas encore encodées en TEI.

L'expérience d'HiSoMA , de l'EHESS et d'AOROC avec les outils du consortium international EpiDoc (Epigraphic Documents in TEI XML) est déterminante. **HiSoMA pourra favoriser la structuration de ce pôle et la généralisation de ses outils.** Les estampages y constituent le document de base des éditions épigraphiques en cours pour les inscriptions de la cité d'Atrax en Thessalie, pour les IGLS et pour le corpus d'inscriptions syllabiques chypriotes. Plusieurs de ces programmes sont engagés dans la production de corpus numériques conformes au modèle EpiDoc (IG Louvre, IGLS, épigraphie de l'Asie du Sud-Est, cachette de Karnak, etc.). **A l'EHESS, le projet ALGreM** prévoit d'éditer un document grec pour chaque siècle, du 5e siècle avant notre ère aux débuts de l'imprimerie, à chaque fois à peu près contemporain de l'auteur et de l'ouvrage qu'il contient ; il étudiera les caractéristiques paléographiques et matérielles, mais aussi l'histoire des pratiques d'écriture, les évolutions du livre comme objet social et les possibilités d'un environnement de publication numérique. Il compte élaborer **un schéma d'encodage** conforme aux propositions de la TEI et du consortium EpiDoc, en utilisant la variété des textes et des contextes pour raffiner celui-ci. Les sources seront avant tout les données ouvertes du projet *Perseus* et le protocole *Canonical Text Services* (CTS) mis en place dans la nouvelle version de cette bibliothèque numérique (http://scaife.perseus.org) sera utilisé.

Toujours dans le cadre d'EpiDoc, en association avec l'AIEGL (Association Internationale d'Epigraphie Grecque et Latine) et son projet européen EAGLE (*Electronic Archive of Greek and Latin Epigraphy*), **AOROC poursuit la mise en place d'une plateforme internationale d'archives épigraphiques électroniques, géoréférencées et partagées**, en partenariat avec la British School at Rome, Oxford University (Charlotte Roueché), l'Institut archéologique allemand (Ph. Von Rummel), le Centre Camille Jullian (UMR8173 CNRS-AMU), l'Association Internationale d'Archéologie Classique (AIAC, E. Fentress), l'Ecole française de Rome (N. Laubry) pour les inscriptions africaines, en collaboration avec l'Université de La Sapienza (S. Orlandi), l'Université de Napoli-L'Orientale (I. Tantillo), l'Académie de Berlin, le musée du Louvre pour ses inscriptions romaines, avec le Musée Lugdunum de Lyon, pour l'ensemble de ses inscriptions, en partenariat avec l'Institut français d'Archéologie orientale pour les inscriptions égyptiennes (hiéroglyphiques, démotiques et coptiques). AOROC inclut dans son champ d'études et de publication de **nombreuses langues rares, notamment italiques ou hispaniques** (projet de base de données en ligne, BEIGE-Base Epigraphique Italie-Gaule-Espagne) ou orientales (syriaque). L'ensemble de ces langues et des nouvelles techniques de publication sur internet font l'objet d'une **formation épigraphique** presque inégalée à l'ENS-PSL et à l'EPHE-PSL, par les membres d'AOROC. Biblissima+ sera en mesure de bénéficier de cette expertise et de la diffuser auprès de la communauté universitaire et scientifique en France et en Europe par le biais de séminaires spécifiques diffusés à distance et en direct par le biais d'un écran interactif de type SpeedTouch UHD 86. Il co-financera avec AOROC et ses partenaires le **système d'information géographique en ligne, Chronocarto, un système ouvert, visant au libre accès de données archéologiques et épigraphiques interopérables**, en permettant leur cartographie à petite, moyenne et grande échelle. Les campagnes d'acquisitions de données se concentreront sur les nombreuses photographies d'archives qui comptent de nombreux inédits par le biais d'un système automatisé de type Slide Strip G1, couplé avec un appareil de photographie numérique à haute définition (de type Nikon D850). Elles seront enrichies des données de terrain au fur et à mesure des différentes missions archéologiques d'AOROC et de ses partenaires. Certains relevés de grandes dimensions et les nombreux estampages conservés par AOROC supposent l'achat d'un scanner A3 supplémentaire. Biblissima+ offrira ainsi une plateforme en ligne de nombreuses sources anciennes en accès libre, valorisant un patrimoine national, presque unique au monde, tout en lui assurant une très grande visibilité internationale grâce à son association avec de prestigieux partenaires européens.

Le **CESCM** a développé un **projet d'édition électronique en XML-TEI soutenu par Huma-Num : cf. TITULUS.** Aux 4500 inscriptions de la France médiévale déjà éditées, s'ajouteront quatre corpus dans les

années à venir. Le Corpus des inscriptions du Royaume latin de Jérusalem (un petit échantillon est déjà en ligne), et plus largement l'ensemble des inscriptions et graffitis en alphabet latin de la Méditerranée orientale 7e-16e s. (environ 2500) font l'objet du projet d'**ERC Starting Grant GRAPH-EAST (Latin as an Alien Script in the Medieval "Latin East", porteur E. Ingrand-Varenne- lauréat en septembre 2020)** . Pour faire le lien avec l'Antiquité tardive, s'ajouteront les inscriptions tardo-antiques et alto-médiévales (les volumes parus du ***Recueil des inscriptions chrétiennes de la Gaule*** **et les nouvelles inscriptions (projet inter-MSH IGAMA, porteur C. Treffort**). Une troisième étape pourra concerner les archives des épigraphistes contemporains (fin 19e-21e s.), qui sont prises en compte dans GRAPH-EAST et IGAMA. La quatrième étape touchera les ressources documentaires, en particulier l'inscripthèque du CESCM, fonds multi-supports spécialisé en épigraphie médiévale. Il regroupe aujourd'hui 400 ouvrages, 40 thèses et mémoires, et plus de 1000 tirés à part, 7000 clichés NB et couleur, papier ou numérique, le fichier de recensement général des inscriptions (25 000 notices), les fichiers formules et biblique, les archives de Robert Favreau et des chercheurs du RICG. Un projet de bibliothèque virtuelle de ce secteur est en cours de réflexion au CESCM et trouvera là toute sa place.

Le corpus s'appuiera sur l'analyse automatique d'images (travaux du cluster 3) pour la transcription des inscriptions. Si les techniques utilisées pour la segmentation automatique en caractères fonctionnent déjà bien pour les manuscrits, cela n'est pas encore le cas pour les inscriptions : détection des lignes, diversité de la documentation, hétérogénéité des supports et des techniques, diversité dans la résolution des images. Il faut établir un protocole d'acquisition d'images et de numérisation. De plus, pour les textes devenus fantômes, illisibles du fait de leur détérioration ou de la superposition de plusieurs couches d'écritures ou d'images, notamment les graffitis, il faut multiplier les analyses de matériaux et les techniques de prise d'images, en utilisant le scan 3D expérimenté dans l'étude de l'art préhistorique et l'archéologie du bâti, mais aussi la photomécanique et les propriétés de la lumière laser développées par l'Institut Pprime (P') (CNRS, UPR 3346, en partenariat avec l'ISAE-ENSMA et l'Université de Poitiers).

Enfin, le **CJM** prévoit l'exploration du **champ totalement délaissé de l'épigraphie latine moderne avec un projet de science participative.** En bénéficiant des compétences en ingénierie pour le traitement informatique de données épigraphiques des autres partenaires (sans doute le CESCM), le CJM prévoit la mise en place d'un site internet et d'un formulaire de saisie permettant à chacun d'adresser, sous forme de texte ou d'image, les inscriptions latines métriques qu'il trouve sur son chemin. Le responsable scientifique contrôle et valide, et la reproduction numérique IIIF ainsi que la notice de l'inscription et son édition en TEI sont disponibles en ligne.

B-I.5.2/ Typologies textuelles du Moyen Âge et de la Renaissance

Implication : CESCM, CESR, CIHAM, CJM, CRAHAM, CRH, HiSoMA, IRHT

Les laboratoires impliqués dans Biblissima+ traitent de typologies textuelles extrêmement diverses.

Le CIHAM a particulièrement développé, y compris avec l'aide de Biblissima, l'édition électronique des **sermons** médiévaux, qui continue d'être l'un des fers de lance du laboratoire. C'est au CIHAM aussi qu'a été élaboré le projet d'édition en TEI de la **glose ordinaire de la Bible (Glos-e)**, redéveloppé depuis par l'IRHT, également largement aidé par Biblissima ; il pourrait être associé à la transcription intégrale et à l'encodage de la Bible d'Olivétan (1535), première bible calviniste (CESR-BVH). Le CIHAM a enfin développé de longue date des compétences dans l'édition électronique de la **documentation pontificale** et des **comptabilités**. A l'ENC-PSL, de nombreux projets sont développés, dont **Miroir des classiques**, déjà présent dans Biblissima. Le CJM développe un projet novateur de chaîne d'étude et d'édition des **documents d'archives** sur le modèle de ce qui existe pour les textes littéraires, en profitant de l'implication dans Biblissima+ des services d'archives. Ces projets permettront d'implémenter DTS (B-I.7.1), de tester et de valider des schémas partagés (B-I.5.3). Le CRAHAM a lui aussi développé un outil d'édition de chartes – l'outil ***e*-Cartae** – qui est utilisé et enrichi au niveau national grâce au financement ANR Actépi. Il s'intéresse aussi à l'édition TEI des actes de la pratique (pouillés, registres, textes de coutumes) ainsi qu'à d'autres types de **textes littéraires** (historiographiques, annalistiques, hagiographiques, liturgiques, poétiques). AOrOc travaille sur l'édition **des textes en hébreu rabbinique**. La base **Jonas** de l'IRHT aimerait développer un corpus textuel en TEI, et le lier au programme **Ricercar** du CESR. **E-Scripta**, comme on l'a vu, inclut bien évidemment l'édition en TEI dans son workflow. A l'IRHT, l'édition en TEI et le repérage de segments de citations par structuration concerne déjà les **encyclopédies** médiévales, les textes glosés, **les textes universitaires et bibliques** – projets qui ont intérêt à travailler avec **Biblindex** d'HiSoMA –, et surtout, en collaboration étroite avec le CRAHAM et le PDN de la MRSH de Caen, l'édition électronique **d'inventaires anciens, avec *Thecae*, *Thesaurus catalogorum electronicus*,** premier corpus au monde de ce type de documents, publié par les Presses Universitaires de Caen. Il s'agit d'inventaires en français, latin, grec, un **corpus des inventaires hébreux** étant en projet lorsque les écritures de droite à gauche seront bien gérées par les logiciels d'encodage TEI (recherches poursuivies par le PDN avec Métopes et Biblissima

d'une part, l'IRHT d'autre part). La collaboration entre l'IRHT et le CRAHAM existe aussi pour l'édition des **encyclopédies médiévales**, ce dernier s'étant spécialisé dans les sources du savoir ichtyologique (programme **Ichtya**).

Biblissima+ va relever un défi particulier grâce à une **collaboration entre latinistes et arabisants** : le développement d'un **environnement de balisage XML-TEI et d'un outil de publication web des textes glosés.** Malgré leur richesse, les corpus de gloses, copiés dans les marges et les interlignes des manuscrits ou sous forme de *glossae collectae*, restent encore largement inédits et de ce fait sous-exploités par les chercheurs. La rareté des éditions critiques concernant ce type de matériel s'explique par les défis singuliers que posent aux philologues la structure des gloses, leurs traditions manuscrites souvent complexes, et le caractère fondamentalement fluide de ces corpus textuels. Pour faire face à ces difficultés, l'édition électronique se révèle, à bien des égards, plus adaptée que le format papier. Mais la publication d'une telle édition suppose de mettre au point des solutions innovantes pour **trouver un équilibre entre les aspects scientifiques et le critère d'ergonomie de l'interface de lecture, sans oublier les possibilités d'interrogation croisée**. L'objectif est double : (1) mettre au point un environnement de balisage XML-TEI propre à l'édition des gloses, utilisable à la fois pour des systèmes d'écritures destroverses ou sinistroverses, (2) développer un visualisateur permettant d'afficher, après balisage, le texte glosé et ses gloses. Le projet est mené par le **CIHAM** en collaboration avec **l'IRHT**, qui mène plusieurs projets dans ce domaine (Gloss-e, Digigloses) et dont le pôle d'édition électronique de sources développe un environnement de balisage des gloses en XML-TEI, ainsi qu'avec le **PDN de Caen et son laboratoire d'édition de sources** (point B-I.5.3). Un travail particulier de mutualisation des schémas d'encodage sera entrepris en concertation avec les porteurs de projets portant sur les textes glosés afin de finaliser les outils d'encodage (environnements de balisage) et de configurer l'outil de publication web des textes glosés (potentiellement fondé sur MaX ou teiPublisher).

A côté de ces projets très raffinés élaborés par les meilleurs spécialistes de telle ou telle typologie textuelle, d'autres projets veulent parvenir à **l'édition de corpus numériquement importants** en utilisant les outils HTR présentés plus haut. Les **BVH** poursuivent le développement d'un corpus de la première modernité en français dans des formats ouverts. Les textes en XML/TEI publiés sur Epistemon sont le résultat de transcriptions quasi-diplomatiques et d'éditions numériques qui fournissent des sources dans leur état original, réutilisables pour différents projets, académiques ou non, en France et hors de France (Canada, États-Unis). Les corpus Rabelais, Montaigne et Ronsard (encore inachevés) permettent l'entraînement des OCR et l'exploitation statistique des données brutes ou annotées, notamment avec l'outil TXM (ENS-Lyon). La reconstitution des bibliothèques privées (fac-similés, éditions numériques, visualisation 2D-3D) suscite beaucoup d'intérêt, pour les auteurs majeurs et par l'exploitation des inventaires indispensables aux recherches sur l'histoire de la diffusion de la lecture et des savoirs. Ce corpus sera augmenté, dans les huit années à venir, d'éditions de textes poétiques manuscrits totalement inédits (projet POMONE), tout en renvoyant à la base de données Jonas (IRHT) et aux référentiels Biblissima. Au sein de ce volet textuel, des projets de philologie numérique et computationnelle ou d'éditions bilingues (français/italien, latin/français, latin/italien) pourront alimenter le développement de nouveaux outils linguistiques et textuels spécifiques. La diffusion des données passe par leur "fairisation", à appliquer de façon plus précise pour les headers TEI par l'attribution d'identifiants pérennes, la compatibilité Dublin Core et IIIF, la mise en place d'un thésaurus générique en coordination avec les groupes « data.cahier » et « typologie textuelle » du consortium Cahier (Huma-Cahier (Huma-Num), qui adapte au domaine textuel l'outil Open-Theso élaboré par la MOM pour l'archéologie, et avec les ontologies élaborées dans d'autres institutions ou disciplines (BnF, ABES, médiévistique, archéologie, etc.).

A l'IRHT, le projet **TELMA-ANACLET** (ANalyse Approfondie de Corpus éLEctroniques Textuels) propose la création d'une chaîne d'édition électronique inversée, mettant l'accent sur les outils et non sur la préparation des données. On y inverse la procédure par rapport à l'édition fine TEI en privilégiant le traitement a posteriori des données par l'utilisateur. Appuyés sur l'IA, les outils d'extraction d'entités nommées et de fouille de texte seront mobilisées, ainsi que des outils de lemmatisation, d'analyse stylistique. Pour ce faire, des interfaces adaptées et ergonomiques seront mises en place. La chaîne d'édition de contenu sera refondue aussi en proposant un espace de publication automatisé simple, déjà mis en place dans la plateforme TELMA1 et dans sa refonte TELMA2, mais qu'il s'agit de développer et de renforcer afin de le rendre aussi simple d'utilisation que l'outil NAKALA de la TGIR Huma-Num. Il s'agit de rendre le contrôle complet de l'édition à l'éditeur, tout comme le contrôle de l'analyse et de l'exploitation revient complètement à l'utilisateur. TELMA – ANACLET deviendrait donc le **lieu de publication de corpus larges et non-enrichis** de Biblissima+.

De toutes ces entreprises d'édition des textes les plus divers en TEI naissent des environnements de travail, proches dans leur principe mais spécialisés et adaptés aux besoins de certaines communautés. Ils ont intérêt à se rapprocher quand ils concernent des typologies textuelles identiques ou voisines, et c'est ce que permettra le cluster TEI de Biblissima+. Ce cluster

favorisera également la création d'autres environnements, permettant ainsi **la généralisation de l'édition en TEI des sources anciennes**, au plus près des spécificités souvent complexes des textes, et donc des besoins des chercheurs. Mais la multiplication de ces environnements et l'enrichissement constant de l'offre en édition TEI fait **naître un autre besoin : une réflexion commune sur cette édition et sur l'annotation des sources anciennes**.

B-I.5.3/ Un « Laboratoire d'édition et d'annotation de sources »

Implication : PDN-MRSH et CRAHAM, CJM, IRHT, HiSoMA et CIHAM

Dans la continuité de la réflexion menée par les membres de Biblissima pour l'édition ou l'annotation en ligne et en open access de corpus de sources (ex : Thecae), on voit donc qu'il faut une **réflexion collective sur les schémas** et qu'il est nécessaire de pousser plus avant la réflexion sur les potentialités qu'offre le numérique, en concevant un laboratoire qui serait un **espace d'expérimentation pour la recherche par la recherche**. C'est la proposition du **PDN de la MRSH de Caen et du CRAHAM**. **Le CJM, HiSoMA et le CIHAM, l'IRHT se joignent à cette proposition**.

Il s'agit de créer un véritable environnement d'édition en TEI :

(1) d'offrir un espace de travail et d'expérimentation collectif et collaboratif qui soit à la fois technique et scientifique pour des éditeurs et annotateurs scientifiques de textes ;

(2) de permettre, selon des modalités variées, la visualisation et la consultation de ces sources dans un portail en ligne ;

(3) de modifier et d'enrichir progressivement le travail produit (*work in progress*).

Il ne s'agit pas d'une plateforme de publication dont les enjeux seraient la validation par les pairs, leur citabilité, leur pérennité. Cependant le travail sur les sources restera protégé par le droit d'auteur.

Ce laboratoire est collaboratif selon deux points de vue :

- collaboratif entre les chercheurs-éditeurs scientifiques et les ingénieurs-chercheurs. C'est à partir d'une réflexion heuristique commune menée au moment même de la conception du programme scientifique, puis tout au long de sa réalisation, que seront modélisés des environnements adaptés aux types de sources éditées, à la démarche scientifique et à son évolution.
- collaboratif entre institutions : il s'agit de prévoir un espace de partage et de discussion (séminaires, ateliers, etc.) afin de se mettre d'accord ou de mesurer les désaccords entre les différentes institutions (PDN, IRHT, ENC, CESR, HiSoMA , CIHAM, CESCM…)

Dans ce contexte d'édition de sources en XML-TEI, on vise :

(1) La mutualisation de schéma(s) XML-TEI pour les projets d'édition scientifique similaires. (Ex : édition critique d'œuvres littéraires anciennes (prose et poésie) ; édition génétique ; correspondance (voir l'expérience et la publication des recommandations au sien du consortium de Cahier) ; édition de chartes (diplomatique) ; édition de compilations de citations ; etc.). Pour cet effort de mutualisation, les équipes caennaises s'appuieront, entre autres, sur des réalisations ou des projets en cours (Projet Ichtya, Actépi, E-Cartae, Chronique de Geoffroi de Malaterra, Édition du Tractatus de Piscibus, etc.) ou en attente de réponse de financement (projet inter-MSH e-CartaeLab). On aboutirait dans l'idéal à un schéma unique commun modulable.

(2) La construction d'outils d'encodage XML-TEI accompagnés de leur documentation *user-friendly*. On pourrait envisager une construction triple :

- pour XXE : construction à Caen grâce à l'expertise du PDN en la matière (les plugins développés seront disponibles en *open source* sur le serveur du PDN depuis l'éditeur XML : http://www.unicaen.fr/recherche/mrsh/document_numerique/seme/) ;
- pour Oxygen : construction à l'IRHT.
- en Open Source : construction au CIHAM, en poursuivant le développement initié avec DARIAH (plugin TEI pour l'éditeur libre JEdit)

(3) un portail de mise en ligne, reposant sur le moteur d'affichage MaX, développé dans le cadre de Biblissima), et permettant :

- d'accéder aux corpus de recherche selon différentes modalités ;
- de construire des outils d'interrogation et de consultation communs.

Ce troisième volet du laboratoire doit permettre de faire **des tests à la fois d'éditions scientifiques (visualiser son corpus, l'éditer, le modifier) et d'outils. On aurait une paillasse virtuelle** sur laquelle on pourrait expérimenter des outils (visualisation de données, consultation de flux parallèles ; graphes ; multi flux de notes ; moteurs de recherches complexes, etc.).

Le laboratoire intégrera également la norme IIIF pour la consultation des images numériques, via le visualiseur mirador, ainsi que le protocole de partage de texte DTS (B-I.7.1).

Le laboratoire sera accompagné d'un **observatoire des pratiques, espace d'échange et de discussion scientifique et technique**, qui débute par la mise à disposition de schémas documentés en usage chez les différents partenaires et qui se poursuivra par des

séminaires que le PDN et le CRAHAM se proposent d'organiser et d'animer. L'objectif étant d'harmoniser aussi bien les schémas de structuration des textes encodés que les environnements de balisage permettant l'annotation.

Dans ce même contexte, le **CJM** pourra proposer ses solutions, outils et schémas afin de créer une **chaîne éditoriale des documents d'archives**, sur la base de la numérisation des documents. Le passage des AN à IIIF, qu'accompagne déjà Biblissima, devrait permettre de mettre en œuvre une telle chaîne en articulant, là comme ailleurs, IIIF et la TEI.

L'IRHT, de son côté, développe un pôle d'édition électronique de sources, dont le but est de trouver un **noyau commun d'enrichissement des éditions de sources en XML-TEI**, et de parvenir avec le PDN de Caen et l'ENC-PSL à une **harmonisation** des environnements de balisage d'éditions critiques capable de gérer non seulement le latin et le grec, mais aussi l'hébreu (voir l'édition en cours du Sefer ha-shorashim, https://shorashim.hypotheses.org) et l'arabe. Les progrès accomplis en matière de transcription automatique des écritures droite-gauche ont ainsi leur pendant en matière d'édition critique.

Le **CIHAM** va poursuivre le développement d'un éditeur XML libre optimisé pour la TEI. Ce développement, très attendu par la communauté des utilisateurs de la TEI, permettra de diversifier – et donc d'assainir – l'écosystème des solutions disponibles en mode *user friendly* pour la pratique de l'encodage de sources en TEI. L'idée est de pouvoir disposer d'une alternative libre, gratuite et viable à Oxygen ou XXE, non pas pour tous les besoins, mais déjà pour l'enseignement et les premiers besoins des projets débutants, qui de toute façon n'utilisent qu'une infime partie des possibilités des éditeurs payants. Les développements feront aussi le pont avec les travaux du PDN de Caen, en prévoyant l'adaptation à jEdit du plugin pluCo. Une première version vient d'être développée, grâce à un financement de DARIAH, sous la forme d'un plugin TEI pour l'éditeur XML libre jEdit. L'angle d'approche privilégié est l'ajout d'un plugin TEI à un éditeur XML déjà existant, ce qui permet d'optimiser l'utilisation des ressources sans réinventer la roue. Le CIHAM voudrait donc :

(1) poursuivre et améliorer le développement du plugin TEI pour jEdit, en prenant en compte les demandes et les retours des utilisateurs;

(2) développer ou améliorer d'autres éditeurs libres pour faciliter leur utilisation avec la TEI, afin de diversifier l'écosystème de solutions disponibles – l'avantage du temps long de Biblissima+ étant de permettre d'adapter les priorités de développement aux évolutions technologiques qui ne manqueront pas de marquer les prochaines années. Ce modèle permettra un développement agile, s'adaptant bien aux besoins de la communauté au fil des 8 années du projet.

On saisit ici toute l'importance d'un **regroupement des acteurs au sein d'un cluster qui sera un lieu de discussion et d'harmonisation des pratiques, et d'invention de nouvelles solutions, afin de fluidifier l'ensemble de la chaîne** et de fournir aux éditeurs matériels, en bout de chaîne (le niveau 6 de l'économie de la donnée : Métopes, OpenEdition / OPERAS), un fichier-maître correctement balisé permettant des publications multimodales. **Cette chaîne de traitement du texte en TEI** intéresse de près **les services de publication** des partenaires qui travaillent en lien avec l'IR Métopes : l'EHESS, l'ENC-PSL, l'EPHE-PSL.

B-I.5.4/ Formations à l'édition numérique en TEI

Implication : HiSoMA et CIHAM, CESR, CRAHAM et PDN-MRSH

L'ambition de généraliser l'emploi de la TEI dans l'édition des sources anciennes les plus variées s'accompagne nécessairement d'un programme de formation, plusieurs des partenaires de Biblissima+ ayant une solide expérience en la matière (PDN et CRAHAM, CESR, HiSoMA et CIHAM).

Le **pôle de Lyon (CIHAM-HiSoMA)** propose de produire et de consolider un ensemble de ressources pour la formation aux nouvelles technologies. Sur le modèle d'un cours DARIAH "Digital Scholarly Editions: Manuscripts, Texts and TEI Encoding" déjà mis en place et exploité expérimentalement lors de plusieurs expériences pédagogiques par le CIHAM en mars et avril 2020, le CIHAM et HiSoMA souhaitent développer d'autres cours, couvrant d'autres besoins en formation des chercheurs de notre communauté de pratique. Concrètement, il s'agit de produire un cours en ligne d'auto-formation et d'organiser des sessions en visio-conférence avec exercices corrigés permettant l'interaction. Le pôle souhaite développer un nouveau cours tous les 2 ans en moyenne, ce qui fait un ensemble de 4 cours sur la durée de Biblissima+. Le CIHAM et HiSoMA seront chacun responsables du pilotage de deux cours tout en collaborant sur la totalité des 4 cours. La collaboration portera sur le cadrage scientifique (choix d'intervenants, sélection des dossiers de candidature voire interventions autour des briques de l'Equipex si c'est pertinent par rapport au thème du cours).

Les 2 cours pilotés par HiSoMA pourront porter sur :

- la modélisation des sources épigraphiques à l'aide du modèle TEI EpiDoc, en particulier autour de la question de la prise en compte de la documentation scientifique (par exemple : estampages).
- La publication de documents TEI et EpiDoc avec l'outil TEI Publisher qui intègre le protocole DTS

et dont une extension pour prendre en compte l'accès aux fragments fait l'objet d'une autre brique du projet.

Le CIHAM a déjà envisagé un cours sur le langage de requêtes XQuery et le système de gestion de bases de données XML natives eXist-db, qui constitue l'un des outils de publication de référence pour les projets en humanités numériques.

D'autres sessions d'accompagnement pourront être organisées à partir des 4 ressources produites, au long de la vie de l'Equipex+ et en fonction des activités mises en œuvre dans le cadre des « environnements Biblissima + » pris en charge par le pôle Lyonnais.

Par ailleurs une ingénieure d'études Lyon 2 d'**HiSoMA** intervient chaque année pour une session d'initiation à la TEI dans le cadre du stage d'ecdotique annuel des Sources Chrétiennes qui forme une quinzaine de jeunes chercheurs et doctorants internationaux à l'ecdotique.

Le **CRAHAM** organise annuellement un séminaire « Erudition & numérique » qui comprend au moins 8 séances. Le séminaire est ouvert, mais il est notamment destiné aux étudiants du Master Histoire d'Unicaen. Il est organisé en partenariat avec le Pôle du Document Numérique de la MRSH. Il permet des échanges sur les pratiques du numérique appliquées aux sources anciennes et médiévales. Il poursuit au moins trois objectifs (1). Un objectif d'information scientifique : faire connaître les projets et les réalisations numériques en cours (dont certains sont menés au CRAHAM en partenariat avec le Pôle Document numérique de la MRSH), ainsi que les apports scientifiques sur lesquels ces réalisations permettent de déboucher. (2). Un objectif d'expérimentation active des outils, des techniques et des réalisations numériques, à partir des problématiques scientifiques disciplinaires qui sont celles du CRAHAM . (3). Un objectif de réflexion épistémologique ouvrant de nouvelles voies de recherche sur la nature même et les finalités des Humanités numériques. L'objectif global de ce séminaire est donc de donner aux étudiants qui le suivent une connaissance suffisante des enjeux, méthodes et outils des Humanités numériques, qu'ils pourront ensuite mobiliser dans leurs propres recherches ou, après le Master, pour leur insertion professionnelle ou la poursuite de recherches doctorales.

Les **sessions de formation à la TEI du CESR** (niveaux initiation et avancé), en collaboration avec le **consortium Cahier et la MSH Val de Loire**, sont également reconnues, et leur développement est intégré à ce programme de formation de Biblissima+, dont le cluster TEI va permettre une harmonisation des initiatives.

B.I.6/ Les défis du patrimoine musical

Implication : CESR, IRHT

Même s'il existe des précédents notables, en particulier l'existence d'une section de musicologie à l'IRHT qui a joué un rôle de premier plan et prépare un projet de recherche et reproduction systématique des fragments musicaux, ou l'initiative *MANNO* à l'EPHE-PSL et à la BnF, ou encore le financement et l'intégration de la base *Comparatio* (IRHT) sur le chant liturgique dans le portail Biblissima, **la dimension sonore en général, et musicale en particulier, est sous-estimée par les historiens de l'écrit**. Or manuscrits et imprimés musicaux participent pleinement du patrimoine écrit du Moyen Âge et de la Renaissance, avec des spécificités qui ouvrent de nouveaux défis à Biblissima+, à commencer par celle d'être des objets matériels conçus pour être réalisés sous forme de performances. Biblissima+, modestement pour commencer, est l'occasion de réintégrer dans un Observatoire des cultures écrites anciennes la dimension sonore et musicale essentielle à ces cultures.

Le programme musicologique Ricercar du CESR entend mettre à la disposition du projet son expertise en matière de traitement numérique de la donnée musicale pour aider à l'intégration du patrimoine musical dans l'étude plus large des cultures anciennes transmises par l'écrit. Les humanités numériques y occupent une place centrale au travers **d'outils de visualisation et d'exploitation de corpus, de reconstitution d'espaces sonores et la constitution d'un fonds documentaire** en lien avec les projets. Outre les publications et les manifestations, le programme Ricercar participe à de nombreux projets de recherche intégralement ou partiellement numériques en collaboration avec des universités françaises et étrangères. Depuis plus de 10 ans le programme Ricercar et Haverford College (Pennsylvanie USA) entretiennent des liens privilégiés autour de projets numériques collaboratifs (DuChemin, Lost Voice, CRIM). Par ailleurs des corpus spécifiques alimentent les bases de données du programme de recherche et entretiennent ainsi des relations entre chercheurs et institutions du monde entier (Music Renaissance in Croatia, Les messes anonymes du XV^{e} siècle, Le corpus de luthistes).

B-I.6.1/ Une cartographie du patrimoine musical et de ses métadonnées

Implication : CESR

Ricercar circonscrit le patrimoine musical écrit du Moyen Âge et de la Renaissance à l'exclusion de l'immense corpus du chant liturgique monodique

chrétien (le plain-chant ou chant « grégorien ») consigné dans les livres liturgiques (avec ou sans notations musicales) qui exigent une approche bibliographique et méthodologique propre, prise en charge en particulier par l'IRHT. Le programme privilégie l'étude des sources proprement musicales. Une particularité bibliographique de l'étude de ces sources est qu'elles ont été intégralement repérées et cataloguées à l'échelle de l'ensemble des bibliothèques mondiales par un programme international initié dans les années 1950 (RISM : Répertoire International des Sources Musicales, http://www.rism.info/home.html). La base de données mise en ligne à ce jour (https://opac.rism.info) fournit les informations bibliographiques de base sur tous les imprimés musicaux produits jusqu'en 1600. Pour les manuscrits, un outil numérique centralisé à Oxford fournit des informations tout aussi exhaustives mais d'une granularité supérieure à celle du RISM, et tient à jour une veille de mise en ligne des sources par les bibliothèques : Digital Image Archive of Medieval Music (https://www.diamm.ac.uk : 4000 manuscrits musicaux jusqu'au 16e siècle). Ricercar tient à jour sa propre veille de sources numérisées en ligne (8000 notices Zotero : la base Jonas de l'IRHT consacrée à la littérature française médiévale souhaite d'ailleurs multiplier les liens vers ces données de Ricercar). Les partitions musicales présentent des spécificités que les modélisations des données bibliographiques doivent intégrer (la mise en page d'une partition est soumise à d'autres contraintes qu'un livre). Dans le contexte de la refonte en cours de ses bases de données, Ricercar a établi des **modèles de données pour la description des sources et des œuvres musicales**. Ses travaux en cours sur les **ontologies de description des œuvres musicales** devraient trouver leur aboutissement dans le cadre de Biblissima+.

B-I.6.2/ Les œuvres : encodage et fouille de la donnée musicale

Implication : CESR

Ricercar dispose d'éditions musicales encodées au **format MEI** (Music Encoding Initiative) et participe activement au **développement de ce standard d'encodage ouvert et scientifique de la notation musicale modelé sur la TEI** (accueil de la conférence annuelle à Tours en 2017). Au-delà de la mise à disposition ouverte d'encodages MEI (Gesualdo online, etc.) permettant de passer de l'image de la partition ancienne à diverses **visualisations** (partition moderne ou autres visualisations graphiques, statiques ou animées, puis audio-numérique, synthétique ou enregistré), plusieurs projets ont développé des problématiques de recherche en matière d'annotation et de fouilles des données musicales, essentiellement pour la **reconnaissance de similarités, avec ou sans intelligence artificielle**. L'objectif d'un **moteur de recherche musical (audio ou écrit)** est l'horizon de recherche de ces travaux, que la recherche académique partage avec les industries culturelles.

B-I.7/ Interopérabilité et analyse des textes

Implication : CJM, HiSoMA, PDN-MRSH, équipe Biblissima+, IRHT

Le dernier grand domaine dans lequel Biblissima+ concentre des forces et souhaite développer les initiatives concerne l'analyse et la fouille de texte, avec un double travail sur l'interopérabilité des corpus et sur la granularité extrêmement fine, permettant d'aller au-delà de ce qui se fait par ailleurs.

B-I.7.1/ Le protocole de partage de textes *Distributed Texts Services* (DTS)

Implication : CJM, HiSoMA, PDN-MRSH

Le projet *Distributed Text Services* (DTS) vise à proposer un standard d'échange et de mise à disposition de textes et de leurs passages via des API spécifiques. **DTS est aux textes ce que IIIF est aux images, d'où son importance extrême**. Il est porté, dans Biblissima+, par les équipes du CJM et d'HiSoMA , en collaboration avec le PDN de la MRSH de Caen.

La spécification des « services de texte distribué » définit une API pour travailler avec des collections textuelles en tant que données gérables par la machine. Elle permet de naviguer dans le contenu des collections textuelles, de naviguer dans un texte unique, de récupérer des textes complets ou partiels, paramètre spécifié pour récupérer des textes ou fragments en XML-TEI. Les paramètres de collecte et de navigation sont spécifiés pour que LD+JSON soit conforme à la norme Hydra du W3C.

Présenté à la conférence TEI 2018 de Tokyo, le standard est aujourd'hui ouvert aux contributions extérieures pour tenter d'assurer sa compatibilité avec le plus grand nombre de projets. Il est fondé sur le partage des textes en TEI, une architecture REST et un catalogue exprimé en JSON/LD. Les partenaires de Biblissima+ sont bien représentés dans l'équipe internationale du projet, avec Thibault Clérice (**ENC-PSL**), Vincent Jolivet (responsable de la mission projets numériques à l'**ENC-PSL**) et Emmanuelle Morlock (**HiSoMA**).Biblissima+ permettra d'accélérer l'utilisation de DTS. DTS sera implémenté dans TEI Publisher, développé par HiSoMA, ainsi que dans les outils éditoriaux du **PDN de la MRSH de Caen.** Le

succès de l'API exige un effort massif en formation des utilisateurs, mais aussi en développement de suites logicielles clients et serveurs (CapiTainS, TEIPublisher, etc.).

B-I.7.2/ Lemmatisation et aide à la traduction des textes anciens

Implication : équipe Biblissima+, HiSoMA, CJM, IRHT

La lemmatisation intéresse de près plusieurs équipes de Biblissima+, à des degrés divers. HiSoMA a créé le prototype de lemmatisation de **Biblindex** (sur un set de 70 000 couples formes/lemmes issus de textes bibliques et patristiques grecs). Elle est parfois à l'état de projet, comme c'est le cas de la base des textes français du Moyen Âge **Jonas** (IRHT), qui voudrait lemmatiser les incipit et explicit de la base pour contourner les aléas orthographiques de l'ancien français.

Biblissima héberge les programmes **Collatinus** et **Eulexis**, dont l'équipex a contribué à créer les versions web : ces outils d'aide à la traduction du latin et du grec se perfectionnent constamment, avec, entre autres, l'accès à des dictionnaires multilingues (langues-cibles : Français, Anglais, Allemand, Espagnol, Italien, Russe, Croate, Portugais) et des fonctionnalités (pour le latin) de scansion, d'accentuation, de statistiques métriques y compris dans la prose, d'analyse automatisée de la syntaxe. Collatinus s'ouvre désormais largement au latin médiéval et aux dictionnaires prosopographiques. **La structure de ces outils serait aisément transposable à d'autres langues**, ce qui va devenir indispensable avec l'ouverture linguistique de Biblissima+.

L'IRHT apporte un corpus lexical lemmatisé de **50 millions de mots de latin médiéva**l (période 800-1200), créé avec le soutien de l'ANR Velum. L'Equipex+ permettra de doubler la taille de ce corpus en y incluant 50 millions de mots issus de textes rédigés entre 700 et 1300, pour l'étude de la diachronie à l'échelle de l'Europe (représentation des textes ibériques ou italiens du 8e siècle ; germaniques ou slaves du 13e siècle) et pour rendre compte à la fois du latin mérovingien et du latin scolastique.

Les référentiels linguistiques sont l'un des points forts du CJM. Son lemmatiseur (**Pie**) est performant, de même que l'application de post-correction (**Pyrrha**), développée au CJM. Des modèles sont aujourd'hui opérationnels en latin classique, en ancien français. Un modèle est en développement pour le français du 17e siècle et pour l'occitan (collaboration avec l'équipe du *Dictionnaire de l'occitan médiéval* à Munich). Des collègues italiens (Padoue) utilisent également Pyrrha et Pie pour la lemmatisation d'un corpus franco-italien. Il faudrait **assurer le chaînage de ces outils avec l'HTR** au sein d'e-Scripta. Le CJM a besoin **d'un service web (schéma, API, application) pour partager les modèles et les ressources (cf. B-I.7.3)**.

B-I.7.3/ Textométrie, stylométrie et alignement

Implication : CJM, HiSoMA, CIHAM

Le CJM dispose déjà, en **dialectométrie / scriptométrie**, d'algorithmes permettant de faire de la classification automatique. Le **Centre de ressources computationnelles pour les langues à variation graphique** qu'il coordonnera se concentrera non seulement sur la question fondamentale de l'annotation linguistique (B-I.7.2), mais aussi sur les traitements qu'elle permet, répondant à des questions omniprésentes dans l'analyse des textes (datation, localisation ; alignement de différentes versions et collation ; détection des entités nommées). L'enjeu est le traitement automatique des langues historiques à forte variation (graphique) et la mise à disposition d'outils (interfaces web, API, algorithmes) et de modèles (essentiellement pour les langues gallo-romanes et le latin). À terme, des services dialectométriques (par exemple, un système de cartes de chaleur) et stylométriques sont envisagés.

Pour la **stylométrie**, des outils très ciblés existent dans *Collatinus* (par exemple l'analyse de la prosodie, du rythme, la recherche de vers ou presque-vers dans la prose). Il serait très stimulant d'avoir des fonctionnalités pour faire automatiquement des rapprochements entre textes en fonction du style ou du contenu, voire de détection des paraphrases, en particulier d'une langue à l'autre.

Le laboratoire HiSoMA a déjà pratiqué plusieurs expérimentations sur des corpus patristiques et médiévaux latins à l'aide d'outils de recherche d'intertextualité, reposant à la fois sur de la lemmatisation et de la textométrie, développés à Göttingen (TRACER) et à Anvers (voir https://www.aclweb.org/anthology/W19-2514.pdf).
Le projet **Biblindex**, riche d'un corpus de 900 000 citations ou allusions bibliques, a vocation à développer ces outils, en les rendant le plus génériques possible pour une application à la recherche de tout phénomène citationnel dans les textes anciens.

A la croisée de l'édition (B-I.5.3), de la lemmatisation (B-I.7.2) et de l'exploitation computationnelle des données textuelles, le CJM et le CIHAM proposent la création d'un outil capable de **reproduire de façon automatisée le processus complet d'établissement du texte**, du niveau macroscopique (alignement par paragraphe ou autre structure textuelle) au niveau microscopique (alignement mot à mot puis établissement d'un apparat typé proposant une

analyse de la proximité des variantes). Un traitement automatique pourra ainsi permettre la classification des variantes graphiques, grammaticales ou sémantiques, voire un classement plus fin s'appuyant sur des représentations sémantiques en fonction du contexte. Du point de vue ecdotique, seront générées des reconstitutions critiques d'archétypes et la visualisation de rapports entre témoins du texte. L'apport de Biblissima+ permettrait d'améliorer le code, de créer et de mettre en production une application web.

Au bout de 5 ans, le chaînage des outils décrits précédemment devrait permettre par exemple, en quelques clics, de sélectionner des manuscrits dans Biblissima+, les passer par l'HTR, et lancer sur eux des analyses de langue, de style, de tradition, faire des rapprochements des écritures via de la vision par ordinateur, etc. On peut imaginer de multiples scénarios de recherche, et donc d'articulation des outils les uns avec les autres. Pour cela, l'ergonomie du portail Biblissima+ jouera un rôle déterminant.

… accélérer le tournant numérique : un portail pour tous …

B-II – Le portail Biblissima et son offre de service

Comment mettre en œuvre un portail Biblissima qui, tout en ayant un périmètre élargi, des fonctionnalités nouvelles, tout en étant connecté avec d'autres laboratoires et projets fournisseurs d'outils, reste simple d'utilisation et capable de répondre aussi bien aux besoins de quelques chercheurs très avancés qu'à ceux de la majorité des usagers de l'ESR et de la Culture, et enfin du grand public, bénéficiaire en bout de chaîne des innovations nées pour répondre à des défis scientifiques ?

B-II.1/ Une offre enrichie mais des interfaces simplifiées

Implication : équipe Biblissima+ et les 7 clusters

B-II.1.1/ Un portail Biblissima plus ergonomique

Implication : équipe Biblissima+

Biblissima, au cours de ses 7 ans d'existence, a créé plusieurs sites web, qui ont épousé et traduit son développement. Au départ, un site Projet **présentant l'équipex**, son organisation, les communautés qu'il fédère, ses initiatives, ses actualités. Ce site-projet comporte des pages de documentation (site Doc), des démos publiées au fil des développements et des innovations (site Démos), et une boîte à outils, Baobab (site Outils), avec un code-couleur facile à comprendre. Quand l'infrastructure de mise en interopérabilité des ressources a été mûre, en avril 2017, a été publié **le site du** portail Biblissima, en version beta, donnant accès aux données accumulées par les partenaires. Le site portail est passé début 2020 de cette version expérimentale à une **version stabilisée**. Au fil du développement du projet s'est affirmée l'importance des référentiels, autour desquels s'agrègent les données des ressources partenaires, une fois celles-ci alignées, désambiguïsées, éventuellement corrigées. A partir de la fin 2018, Biblissima a commencé à publier ces référentiels patiemment constitués, sur une plateforme qui est un outil de plus, facilement moissonnable par les machines, data.biblissima, qui ne cesse de s'enrichir (en juillet 2020, 228 363 entités publiées, soit plus de 9 millions de triplets RDF dans le triplestore ; 245 564 entités prévues à la rentrée académique 2020). Elle est l'épine dorsale du portail Biblissima, mais aussi d'un autre site ouvert pour répondre à la demande internationale d'intégration à Biblissima : IIIF Collections of Manuscripts and Rare Books. Cette plateforme permet d'interroger sur n'importe quelle entité les collections numérisées interopérables de diverses bibliothèques françaises et étrangères (en juillet 2020, 73 000 manuscrits environ), de façon extrêmement simple et rapide, mais sans les faire entrer en interopérabilité profonde comme c'est le cas des ressources Biblissima. 65% des visiteurs uniques de l'ensemble des sites viennent de France ; 21% des visites étrangères viennent des Etats-Unis. Depuis quelques mois, la home page https://biblissima.fr facilite l'accès aux 4 sites les plus visités.

Aujourd'hui, nous constatons plusieurs phénomènes. **Le portail Biblissima reçoit de plus en plus de visites** (32 519 visites en 2019, +81% entre 2018 et 2019, dont plus de 50% viennent de l'étranger) ; la mise en

interopérabilité de la base iconographique *Initiale*, largement relayée sur Twitter, va amener de nouveaux usagers travaillant non seulement sur les textes mais aussi sur les images. En parallèle, le site projet attire moins les internautes : c'est normal, puisque le projet est réalisé. **La boîte à outils connaît un grand succès** (52 915 visites en 2019 ; l'outil Collatinus-web a reçu 394 669 visites entre 2016 et 2020 ; de mars à mai 2020 par suite du confinement, + 84% pour Collatinus, soit + 136,8% par rapport à la même période en 2019 ; + 75% pour Eulexis, soit + 104% par rapport à la même période en 2019). Le site **IIIF Collections** a pour le moment moins de visites (6 245 visiteurs uniques en un an), mais il fait plus le buzz et 84% des visites viennent de l'étranger. C'est lui qui donne le plus de place aux initiatives internationales, et les bibliothèques qui y sont représentées pensent faire partie de Biblissima, ce qui n'est pas tout à fait vrai en réalité, mais prouve l'existence d'une demande, d'une attente. Encore plus simple d'usage que Biblissima, il plaît pour cette raison, et aussi parce qu'il donne **accès potentiellement aux manuscrits du monde entier**. L'intégration de ces fonds numérisés grâce à une interopérabilité relativement légère peut être assez rapide, et elle alimente les référentiels Biblissima, en particulier celui des cotes. Elle mérite donc d'être développée.

Toutes ces observations nous portent aux conclusions suivantes.

Le site Projet n'a plus la même actualité, il doit être pratiquement archivé. Le **site Doc** garde son utilité, en particulier pour des partenaires candidats à une interopérabilité profonde, ou qui veulent comprendre les technologies mises en oeuvre. C'est le cas en particulier pour le protocole d'interopérabilité des images IIIF, mais aussi pour l'ontologie et les thesauri. Le portail doit, comme aujourd'hui, **fournir le mode d'emploi pour rejoindre ses ressources**.

Il faut **un seul portail qui combine IIIF Collections et Biblissima** parce que c'est le meilleur moyen, relativement peu coûteux, de donner à cette infrastructure nationale un rayonnement international. Du point de vue de la recherche, cela s'impose : intuitivement, les internautes y cherchent les manuscrits du monde entier, pour les visualiser et les comparer à d'autres, les replacer dans des ensembles signifiants. **Ils s'attendent à une exhaustivité pour la France, et rêvent d'une tentative d'exhaustivité pour le reste du monde**. Grâce à **une interface très simple**, ils veulent chercher leurs objets (cote, nom de lieu, de personne, marque de provenance, thème iconographique, oeuvre…) dans l'ensemble de ce qui est offert sur la toile, mais avec des résultats de qualité, qui ne les encombrent pas de pseudo-résultats et de pistes inutiles.

La mission du portail Biblissima+ sera de continuer d'accroître l'agrégation de données, le partage et la réutilisabilité de ces données. De nouveaux axes feront l'objet de développements spécifiques :

- la mise à disposition de la bibliographie concernant les entités indexées dans le portail,
- le référencement systématique des éditions électroniques de textes (réalisées ou non au sein du réseau de partenaires),
- la mise en place de connecteurs vers les différents sites des projets partenaires (science des matériaux, plateformes de transcription et d'édition, *corpora* textuels bilingues) et de mécanismes de récupération des annotations IIIF qui seront faites par les partenaires, en particulier dans les projets de transcription ou de reconnaissance automatique de formes et d'écritures manuscrites.

La recherche de simplicité doit gouverner la façon dont se mettront en place les liens évoqués plus haut avec les autres ressources scientifiques et la façon dont ils seront accessibles : **de simples « boutons »** pour la bibliographie, les archives numérisées, les projets de recherche seraient sans doute la meilleure solution, afin de **ne pas encombrer l'interface**.

Il faut également travailler sur **la simplicité et l'ergonomie de la présentation des résultats, et de leur export** pour une réutilisation, en tenant compte des niveaux variables de compétences des usagers : de simples pdf, de simples listes ou fichiers csv pour ceux qui le souhaitent, un sparql endpoint et des exports rdf pour les usagers plus avancés, des visualisations plus ou moins ludiques pour ceux qui en ont besoin (usages pédagogiques, journalistiques, vulgarisation, sites web etc.), l'introduction dans le portail de la dimension sonore, pour les manuscrits musicaux mais aussi pour les textes.

Dans la présentation des résultats, il faut surtout veiller à l'ergonomie de la **constitution des corpus de données, et à la possibilité de les retravailler et de les partager**.

B-II.1.2/ La gestion du bouquet d'outils

Implication : équipe Biblissima+ et les coordinateurs des 7 clusters

A partir de n'importe quel set de données, le portail Biblissima doit permettre **d'ouvrir son corpus directement dans un environnement de travail choisi, ou de chaîner à sa convenance les outils pertinents pour ce que l'on veut faire.** Il y a ici un enjeu très important d'ergonomie et donc de graphisme, de souplesse, d'adaptabilité aux besoins, qui se combine avec une bonne information de l'utilisateur sur les possibilités offertes par tel ou tel outil. La boîte à outils Baobab pourrait être entièrement repensée pour cela.

Une telle interface ne s'improvise pas. **Sa conception se fondera sur la réflexion des clusters et la mise en commun des besoins, des solutions imaginées, pour assurer la fluidité du chaînage du point de vue de l'utilisateur**. En un premier temps, le portail Biblissima créera de simples boutons orientant vers les sites des partenaires du bouquet, mais cela ne peut être qu'une solution d'attente.

B-II.2/ Accélérer le tournant numérique des communautés : nouveaux espaces virtuels de recherche et de formation

Implication : équipe Biblissima+, GED

Le portail doit enfin fournir des outils de formation, afin d'accélérer l'acculturation des communautés et de créer de nouveaux usages dès les premiers contacts avec la recehrche sur les cultures écrites anciennes.

Au niveau du portail lui-même comme au niveau des environnements conçus par les partenaires du bouquet Biblissima+, il faut accroître la possibilité de partager des sessions, et de créer de **nouveaux espaces virtuels de travail et de formation** à partir des sets de données en ligne. La situation de confinement qu'a connue le monde au printemps 2020 a montré l'extrême importance de cette possibilité.

On peut pousser cela jusqu'à la **création d'une bibliothèque / salle de classe en 3D, partageable en présentiel et à distance**, qui pourrait également connaître des usages pour la **médiation scientifique** : par exemple la reconstitution d'une ou plusieurs bibliothèques médiévales (quand c'est possible) à partir des inventaires anciens édités en TEI dans la collection Thecae, créée par Biblissima aux Presses Universitaires de Caen, ou la possibilité de compulser virtuellement et collectivement, dans la salle de cours ou de séminaire, la bibliographie *ad hoc*, ou encore de manipuler un objet ancien porteur de texte et d'image en 3D. Au niveau de l'Île-de-France, en lien avec le CPER, le GED et les bibliothèques de Plaine-Commune, cela pourrait être réalisé grâce à une collaboration avec le LUTIN par ex. (EPHE-PSL, P8, Cité des sciences), tout en mettant à profit l'expérience de la MRSH de Caen (salle d'immersion 3D). Une fois ce type de réalisation créé, on voit très facilement comment l'utiliser dans d'autres contextes scientifiques, avec des objets tout différents.

❦

L'infrastructure dans son ensemble constitue ainsi une offre de service à destination des producteurs de données que sont les bibliothèques, les services d'archives, les équipes de recherche, les étudiants, les citoyens en général, en favorisant leur cheminement vers l'interopérabilité et le partage des données. De même, elle est tout entière une offre de service pour les particuliers, à qui elle offre un accès unique et simplifié à des sets de données interopérables, réutilisables, et un espace-atelier pour les retravailler, les enrichir, et les remettre à disposition de tous.

Elle joue le même rôle pour les outils développés au fil du programme par elle-même et par les bénéficiaires des financements des projets partenariaux. Elle inscrit cette **fonction d'opérateur d'écosystème** dans le projet plus global du Campus Condorcet, en lien avec la TGIR Huma-Num qui peut récupérer l'ensemble de ses productions pour enrichir son offre de service, et en région grâce aux relations des équipes avec les projets de CPER et les MSH.

Son offre de service est enfin incarnée par l'équipe portail Biblissima+, son rôle de coordination et de valorisation de l'action des clusters, et d'accompagnement en ligne et en présentiel de communautés en pleine évolution.

Abréviations :

CERL : Consortium of European Research Libraries

CPER : Contrat de Plan Etat-Région

DRAC : Direction Régionale des Affaires Culturelles

ERIC : *European Research Infrastructure* Consortium

FAIR : *Findable*, *Accessible*, *Interoperable*, *Re-useable*

FNE : fichier National d'Entités

HTR : *Handwritten Text Recognition*

IIIF : *International Image Interoperability Framework*

IR : infrastructure de recherche

ISMI : *International Standard Manuscript Identifier*

MEI : *Music Encoding Initiative*

MESRI : Ministère de l'enseignement supérieur, de la recherche et de l'innovation

TEI : *Text Encoding Initiative*

TGIR : Très Grande Infrastructure de Recherche

XML : *Extensible Markup Language*

Biblissima

Patrimoine écrit du Moyen Âge
et de la Renaissance

Biblissima bénéficie d'une aide de l'Etat gérée par l'ANR au titre du programme « Investissements d'avenir », portant la référence ANR-11-EQPX-0007