Programme de numérisation enrichie de CollEx-Persée

Axe « Numérisation concertée de la littérature scientifique »

Présentation de la démarche

Conformément à la feuille de route de préfiguration de CollEx-Persée 2, les grands axes du futur programme de numérisation ont donné lieu en 2023 à un travail d'approfondissement et de structuration. Un groupe de travail a été mis en place pour préparer l'axe dédié à la numérisation concertée de la littérature scientifique. Ce groupe était ouvert

- à l'ensemble des bibliothèques et opérateurs partenaires du programme de numérisation concertée entre Persée et les Plans de conservation partagée des périodiques thématiques nationaux, mené de 2020 à 2023 dans le cadre de CollEx-Persée 1
- aux bibliothèques ayant soumis à l'AMI de 2022 des projets faisant place à des opérations de numérisation de la littérature scientifique

20 bibliothèques ou opérateurs ont été représentés lors de la réunion de lancement de la démarche et 14 ont contribué activement aux groupes de travail mis en place à l'issue de cette réunion.

Groupe 1 : Diffusion et valorisation des différents acteurs

Grégory Colcanap, copilote	Bib Cujas
Gabrielle Richard, copilote	Persée
Marc Martinez	Bib. SciencesPo
Lucie Albaret	SCD U. Grenoble-Alpes
Emilie Barthet	SCD U. Bourgogne
Julie Mistral	Abes
Cécile Cérède	Bib Cujas
Olesea Dubois	Bib. SciencesPo
Faizah Mokhtari	SCD AMU

Groupe 2 : Priorités, critères de choix des corpus et indicateurs

Marc Martinez, copilote	Bib. SciencesPo
Delphine Coudrin, copilote	SCD U. Bordeaux Montaigne
Julien Baudry, copilote	SCD U. Bordeaux Montaigne
Gabrielle Richard, copilote	Persée
Lucie Albaret	SCD U. Grenoble-Alpes
Emilie Barthet	SCD U. Bourgogne
Aleth Tisseau des Escotais	Bib de l'Observatoire
Alix Merat	Bib Cujas
Cécile Cérède	Bib Cujas
Olesea Dubois	Bib. SciencesPo
Sophie Ientile	La Contemporaine
Lydie Ducolomb	SCD U. Lyon 1
Juliette Jestaz	BIS

Groupe 3 : Méthodologie et consultation des communautés de recherche

Alexandra Gottely, copilote	Bib Cujas
Olesea Dubois, copilote	Bib SciencesPo
Marie-Madeleine Géroudet, copilote	SCD U. Lille
Gabrielle Richard, copilote	Persée
Marc Martinez	Bib SciencesPo

Lucie Albaret	SCD U. Grenoble-Alpes
Emilie Barthet	SCD U. Bourgogne
Cécile Cérède	Bib Cujas
Sylvie Lemaire	La Contemporaine
Benjamin Guichard	BULAC
Sébastien Dalmon	BIS

Des séances de travail programmées d'avril à juillet, ainsi qu'un échange avec les directions de l'Abes et de l'INIST, ont abouti au cadrage du plan, présenté dans ce document, concernant

- le périmètre du plan
- la valorisation et la dissémination des données produites.

L'organisation opérationnelle et la mobilisation des communautés de recherche feront l'objet d'une poursuite des travaux en 2024.

Périmètre de la numérisation concertée

Définition des priorités de numérisation

1. Numériser les publications scientifiques

Le plan de numérisation concertée porte sur les publications scientifiques. Mais **le périmètre doit être entendu au sens large** et inclure par exemple les bulletins de sociétés savantes ou des revues dont la dimension scientifique est discutable mais qui appartiennent au champ scientifique parce que des chercheurs y publient.

L'usage doit donc guider les choix : il faut **apprécier la contribution d'une publication au développement d'un champ disciplinaire**. Ce sont les communautés scientifiques qui peuvent dire si une publication est importante dans leur champ.

Si la numérisation des sources n'est pas l'objet du plan de numérisation concertée, la définition des corpus prioritaires devrait **tenir compte de l'existence de projets complémentaires de numérisation de sources** (ex. *Bulletin de la société française de philosophie* et correspondance des sociétaires).

2. Donner la priorité aux périodiques

La numérisation concertée doit porter **en priorité sur les périodiques**. C'est le type de document le moins bien signalé dans les catalogues et qui **bénéficie donc le plus du référencement** permis par la numérisation telle que mise en œuvre par Persée (métadonnées au niveau de chaque article). Par ailleurs, ce sont des collections qui présentent **de forts enjeux de conservation** (place occupée, complétude et état des collections, communication particulièrement dans le cas d'un stockage délocalisé).

Pour la même raison de signalement, il est intéressant d'inclure dans le plan **les ouvrages collectifs** (colloques notamment). Mais ils représentent un ensemble beaucoup plus réduit et plus difficile à identifier. Pour les années 60 et 70, la frontière avec la littérature grise est parfois floue, les actes étant souvent publiés sous forme dactylographiée.

Le deuxième ensemble identifié comme prioritaire pour un plan de numérisation concertée est celui **des thèses**. Différents programmes de numérisation, menés souvent en concertation avec la BnF, ont pu assurer une certaine accessibilité aux thèses du XIXème siècle mais le travail reste largement à faire pour le XXème siècle. La numérisation des thèses doit permettre de répondre à deux ambitions majeures : assurer la disponibilité et la mise à disposition de documents qui sont **souvent des unica**, répondre aux besoins de la recherche actuelle dans des secteurs ou sur des **thèmes émergents** pour lesquels les publications scientifiques sont moins nombreuses. Le caractère massif de ce corpus, l'objection constituée par l'inégal intérêt scientifique des thèses et la difficulté liée à l'obtention des droits appellent deux préconisations:

- Opérer une sélection non pas titre à titre mais par ensembles, définis de façon thématique en fonction des thèmes émergents et de statistiques de consultation, en conservant une approche sérielle.
- Numériser en priorité les thèses dont la diffusion en accès ouvert sera possible.

La **littérature grise** n'est pas en tant que telle une priorité du plan. Néanmoins, la **volonté de traiter les séries de façon exhaustive** doit amener à inclure dans le plan la numérisation des débuts de collection, qui peuvent correspondre à des publications plus rares et moins normées, s'apparentant à de la littérature grise.

Le cas des **publications dont l'intérêt scientifique tient aux données incluses** dans la publication doit être examiné de façon spécifique, par les chercheurs du domaine. Ces publications, tout comme les **éditions de sources ou de textes**, peuvent demander une autre forme d'accès que celui offert par Persée (accès en base de données notamment). Ces gisements appellent un important travail d'identification et d'évaluation de la pertinence des formats produits. Ils ne pourraient donc être pris en compte dans un programme concerté que dans un deuxième temps.

3. Numériser des contenus sous droit

Les programmes de numérisation mis en œuvre par les bibliothèques universitaires ou par la BnF ont massivement porté sur la période libre de droit et spécifiquement sur le XIXème siècle. La plus-value du plan consiste donc à numériser les publications du XXème siècle, y compris les publications sous droit, en organisant ou prenant en charge la collecte des droits, ainsi que le fait Persée pour le développement de son portail de revues.

Des attentes particulières, attestées par la consultation des collections imprimées, portent sur les années 60, 70 et 80. Des statistiques de consultation des collections imprimées pourront étayer cette observation.

4. Quelle place pour les publications étrangères ?

Exclure du plan la numérisation de publications étrangères risque d'exclure de fait certaines disciplines (et notamment les sciences). Par ailleurs, la disponibilité des publications numériques étrangères est très inégale selon les aires géographiques ou linguistiques. Les publications anglophones sont assez largement disponibles, ce qui n'est pas le cas des publications d'Europe de l'Est ou du Maghreb. Par ailleurs, certains corpus très spécifiques et rares, mais importants dans leur domaine, de publications scientifiques étrangères ne semblent pas devoir être écartés du programme. En outre, il semble intéressant d'étendre le plan aux revues étrangères consacrées aux études françaises.

Néanmoins la numérisation de publications étrangères rencontre deux difficultés majeures : la concurrence possible avec d'autres projets menés hors de France, la difficulté de recueillir les droits à l'étranger.

Pour cette raison, on estime qu'une publication étrangère pourra être prise en compte aux conditions qu'il y ait une demande explicite des chercheurs et qu'on ait instruit et confirmé la question de sa non-numérisation dans son pays d'origine.

L'apport de CollEx-Persée pourra consister en priorité en :

- Un conseil juridique pour la numérisation de publications étrangères,
- Le cas échéant un cadre de partenariat et de négociation avec des acteurs internationaux pour la numérisation de publications étrangères.

Complémentarités et partenariats

1. Faut-il exclure de re-numériser des contenus déjà disponibles ?

Plusieurs situations peuvent conduire à envisager une re-numérisation : le besoin de produire des données de meilleure qualité ou de nouveaux formats, le besoin de produire un corpus homogène quand une partie est déjà disponible ailleurs, le souhait de rendre librement accessibles des ressources soumises à abonnement.

- La priorité du plan doit être la mise en ligne de nouveaux contenus.
- Il faut éviter des numérisations complémentaires qui auraient pour conséquence de disperser les publications et privilégier les projets portant sur de longs segments continus.

- Des questions de qualité des données peuvent rendre nécessaire une re-numérisation mais ces cas doivent être arbitrés individuellement et dans un souci d'économie de moyens.
- Pour garantir un accès à l'ensemble d'une publication et éviter toute redondance dans la numérisation, une interopérabilité entre Persée et Gallica est à construire, pour laquelle CollEx-Persée constitue un cadre de discussion.
- Pour les contenus déjà disponibles sur des plateformes payantes, la numérisation pour diffusion ouverte n'est pas dans les objectifs du plan.

2. Quel équilibre établir entre les ressources de niche et les ressources largement partagées ?

On considère ici une ressource de niche comme une ressource dont l'audience est étroite en raison d'une forte spécialisation disciplinaire ou géographique, mais qui est une référence importante dans son domaine.

Si l'importance des usages et de l'impact suffisent à justifier la numérisation de publications très présentes dans les collections des bibliothèques, le plan ne doit pas exclure les ressources plus spécialisées et de moins large audience. Il est notamment intéressant de pouvoir y intégrer des revues savantes locales. Dans le cas de ces publications, la numérisation permet tout spécialement de répondre aux enjeux suivants :

- La fréquente dispersion des collections pour ces titres ;
- La fragilité des modèles économiques de ces publications ;
- Leur spécialisation qui correspond à des recherches pointues et reflète des modalités spécifiques de constitution du savoir scientifique ;
- Le besoin exprimé par les communautés scientifiques de disposer de ressources rares.

- Le plan doit inclure des ressources très spécialisées, choisies avec les communautés de recherche, à côté de publications de plus large audience.

3. Quelle articulation créer avec les objectifs de conservation ?

Dans le programme mené de 2020 à 2023, la difficulté à obtenir une collection massicotable a conduit à imaginer de relier, pour en faire une collection de conservation, la collection numérisée par Persée. Cela permettrait de consacrer à la numérisation une collection en bon état et d'avoir une collection de conservation ayant fait l'objet d'un récolement page à page.

- Si elle répond à une très grande partie des usages, la numérisation aux standards de Persée ne se substitue pas à la consultation du papier dès lors que la recherche porte sur des particularités d'exemplaires ou des contenus non scientifiques (publicités par exemple) qui ne sont pas diffusés sur le portail. Numérisation et accessibilité d'une collection de conservation sont donc des objectifs complémentaires.
- Relier la collection numérisée s'inscrit dans les objectifs du plan dès lors qu'il n'existe pas déjà de collection de conservation identifiée (préconisation du nombre de collections de conservation à préciser) et que le titre n'est pas très largement présent dans les collections de l'ESR.
- Dans le cas où une collection de conservation existe déjà, la recherche d'une collection massicotable devrait être élargie à des institutions publiques ou privées extérieures aux PCP et à l'ESR qui n'ont pas de mission de conservation.
- Une bonne articulation du programme de numérisation avec les enjeux de conservation partagée demande d'établir des préconisations relatives au nombre de collections à conserver à l'échelle de l'ESR.
- La numérisation doit être considérée dès la mise en place des plans de conservation partagée et à défaut au plus tôt dans la planification des opérations de gestion des collections pour

optimiser le lien entre numérisation, désherbage et constitution d'une collection de conservation.

- Un outil permettant de visualiser les collections de Persée avec celles des plans de conservation paraît indispensable.
- Le métrage linéaire des collections numérisées par Persée et les statistiques de consultation de ces titres sur le portail sont des indicateurs pertinents pour les établissements qui sont partie prenante au plan.

Proposition d'application et pondération des critères

Plusieurs principes de base ont émergé pour aider à l'application des critères :

- La méthodologie doit être appliquée par champ disciplinaire, pour tenir compte des spécificités éditoriales de chaque discipline et pour s'appuyer autant que possible sur l'existence et l'expertise des Plans de Conservation Partagée (PCP). Dans la mesure du possible, l'ensemble des champs disciplinaires seront représentés dans le plan.
- L'usage et l'importance de la publication dans son champ disciplinaire constituent un critère majeur pour sa sélection. Leur appréciation repose sur des données chiffrées (statistiques de consultation, de prêt, de PEB, etc.) mais également sur une expertise et une connaissance du domaine éditorial.
- Le rôle de la communauté des chercheurs pour évaluer l'intérêt scientifique d'une publication est central, sa mobilisation et les difficultés qu'elle pose doivent faire l'objet d'une réflexion de méthode qui accompagnera la mise en œuvre du plan.
- Les enjeux de conservation n'interviennent que de façon complémentaire aux autres critères.

Table des critères

Critère	Mode d'application du critère	Priorité du critère
Intérêt scientifique	Définition large de « scientifique » : où des chercheurs publient ou qu'ils citent. Est mesuré par l'appréciation de la contribution de la publication au développement d'un champ disciplinaire. Est évaluée à la fois par les bibliothécaires et les chercheurs Prend en compte l'existence de projets complémentaires de numérisation de sources	Elevé : plus l'intérêt scientifique est avéré, plus la publication est susceptible d'être numérisée.
Type documentaire	Déterminer le type de la publication concernée.	En fonction du type, priorité variable, avec comme ordre : 1. Périodiques ; 2. Ouvrages collectifs ; 3. Thèses ; 4. Littérature grise dès lors qu'il y a publication en série ; 5. Publications/éditions de sources ou données
Statut juridique	Examen du statut juridique de la publication : dans le domaine public, sous droits, l'éditeur existe-	Le statut juridique ne doit pas être un critère de priorisation ou

	t-il encore ?, l'éditeur publie-t-il encore ?, etc.	de rejet a priori. La collecte des droits peut être prise en charge.
Période de publication	Dates de la publication	Le XXe siècle est à privilégier, notamment les décennies 60, 70, 80.
Nationalité	Nationalité de l'éditeur	Priorité donnée aux publications françaises, mais n'est pas un critère excluant : une publication étrangère peut être numérisée s'il y a demande et après examen du statut éditorial.
Contenus déjà numérisés	La publication est-elle déjà numérisée ? Si oui, sous quel format, qualité ou exhaustivité ?	Priorité donnée aux contenus inédits en version numérique. On privilégie des segments continus non encore numérisés. La numérisation d'un contenu déjà numérisé est arbitrée individuellement et justifiée par d'autres critères.
Audience	La publication concerne-t-elle un domaine large ou spécialisé ? Quelle est l'audience potentielle de la publication ? Est-elle très présente dans les bibliothèques ?	Le critère de l'audience n'est pas excluant pour une « ressource de niche » dès lors que la communauté spécifiquement visée la considère comme une ressource de base pour la discipline.
Enjeux de conservation	Une collection de conservation est-elle identifiée et accessible ? Peut-on trouver facilement une collection massicotable ?	Les enjeux de conservation ne sont pas prioritaires. Numérisation et conservation de l'imprimé sont envisagées comme complémentaires. La nécessité de re-relier une collection massicotée ne doit pas être bloquante.

Enrichissement, dissémination et valorisation des corpus produits

Au-delà de l'affichage institutionnel, c'est l'adéquation aux usages qui est l'objectif principal de la diffusion et de la valorisation des contenus. Deux types d'usages sont visés : la consultation des corpus numérisés et la fouille à des fins de recherche. Les enjeux de dissémination sont très étroitement liés à ceux d'enrichissement : inscrire les métadonnées dans un écosystème large (autres bases bibliographiques, référentiels, outils d'indexation et de visualisation, etc.) permet à la fois de les consolider et d'en augmenter la découvrabilité.

Signaler les ressources numérisées et favoriser leur consultation

La numérisation augmente considérablement la visibilité des publications en séries, dont le catalogage classique ne permet pas de signaler finement les contenus. Optimiser l'indexation et disséminer les métadonnées produites sont donc deux dimensions essentielles du plan de numérisation concertée, qui portera largement sur des périodiques.

- La diversité des usages invite à **privilégier la dissémination** des métadonnées au modèle d'un portail fédérateur. C'est aussi le moyen d'inscrire les ressources numérisées au sein de différents projets d'éditorialisation ou de mise en valeur sans recourir à une double diffusion.
- L'interopérabilité des différents systèmes à travers lesquels sont propagées les métadonnées (plateforme de Persée, outils de l'Abes, outils de découverte des bibliothèques, entrepôts bibliographiques internationaux, etc.) ne doit pas être envisagée de façon seulement abstraite, mais considérée de façon pragmatique en prenant en compte les différents profils d'implémentation FAIR des différentes plateformes. A partir de scénarios privilégiés fondés sur les usages des chercheurs, **des bonnes pratiques pour la dissémination des métadonnées** du programme doivent être établies. Au-delà du plan de numérisation concertée et des contenus produits pour le portail Persée, l'ambition serait de conforter une voie d'alimentation des SIGB et outils de découverte basée sur des données produites par l'ESR. Ce corpus pourrait constituer un pilote pour la démonstration d'un écosystème vertueux garantissant souveraineté, qualité et visibilité à l'international.
- La numérisation des publications scientifiques produit des documents numériques qui ont vocation à s'inscrire à la fois dans l'écosystème de la publication numérique native et dans celui des ressources numériques patrimoniales. Ceci demande **d'articuler différents standards** et peut faire émerger la nécessité d'une évolution des normes. Il demande au minimum de définir les cibles et les bonnes pratiques de dissémination à travers ces deux écosystèmes.

Ces différents objectifs rencontrent pleinement le projet d'établissement 2024-2028 de l'Abes et le plan de numérisation concertée peut fournir un cas d'usage central pour les réflexions de l'agence concernant les flux de métadonnées pour la documentation électronique. Plusieurs critères de priorisation sont en effet communs au plan de numérisation concertée et aux cas d'usage privilégiés par l'Abes (cf § 3.1 du projet d'établissement : https://projet2024.abes.fr/docs/2.4/projet2024)

Rendre les corpus produits disponibles à la fouille

Le caractère manipulable des corpus et la possibilité de les rapprocher de corpus complémentaires déterminent leur disponibilité effective pour des opérations de fouille.

- **Verser le corpus produit dans ISTEX**, à des fins de fouille et non à des fins de consultation, pour que le TDM en soit facilité grâce à la constitution d'un corpus rendu homogène par des pré-traitements. Ce versement permettrait de venir alimenter un gisement déjà massif et apparenté, en renforçant la présence des publications scientifiques françaises et francophones. L'INIST est favorable à cet objectif, qui demande une évolution technologique d'ISTEX en cours d'implémentation (pour OpenEdition).
- Afin d'**identifier les formats et données utiles** (texte brut, TEI, métadonnées, plein texte, etc.) et de cibler les efforts faits pour l'amélioration de leur qualité, il est nécessaire de s'appuyer sur des cas d'usage.

Enrichissements supplémentaires

La structuration fine des contenus et l'alignement des auteurs sur les principaux référentiels sont les deux enrichissements majeurs apportés aux documents numérisés sur la plateforme de Persée.

Parmi les nombreux enrichissements qui peuvent être souhaités pour valoriser les ressources produites, **l'exploitation des citations est une priorité**. Les liens de citation constituent un point d'entrée et un moyen de navigation qui augmentent massivement l'accès au corpus dans une logique de bibliographie scientifique. Ils répondent aussi à des questions de représentation des réseaux à l'œuvre dans la constitution du savoir et des sciences. L'intérêt de cet enrichissement est transversal à l'ensemble des publications scientifiques traitées. En revanche, les outils d'identification des citations peuvent avoir une efficacité variable selon les périodes et les disciplines, qui déterminent différents formalismes. Actuellement, la plateforme Persée permet l'identification des citations internes au corpus de Persée. Le lien vers des bases externes ne pourra se faire qu'au moyen d'une plus grande automatisation.

- Un outil pour **l'identification des citations et leur alignement sur des bases bibliographiques** est à développer/entraîner pour l'enrichissement du corpus produit. L'INIST est prêt à s'associer à plusieurs étapes d'un tel projet : test des outils aujourd'hui utilisés dans ISTEX pour la reconnaissance et structuration des références bibliographiques sur le corpus de Persée, réintégration des enrichissements dans le système d'information de Persée.
- Les outils utilisés pour l'enrichissement du corpus du plan de numérisation concertée doivent être **rendus disponibles comme service** pour les données produites sur la plateforme de Persée.

Valorisation des corpus enrichis

Le plan de numérisation concertée doit avoir non seulement pour objectif de répondre aux usages émergents de fouille de texte, mais également d'en favoriser l'essor et de contribuer au développement des compétences qui y sont nécessaires, aussi bien du côté des chercheurs que des personnels de bibliothèques.

- Des corpus d'entraînement et des démonstrateurs doivent être proposés pour susciter de nouveaux usages des ressources produites. L'accompagnement des chercheurs se décline en plusieurs étapes dans lesquelles les bibliothèques ont un rôle à jouer : expliquer la nature des données mises à disposition (métadonnées, référentiels, etc.), illustrer les exploitations

possibles, donner accès à des outils pré-paramétrés, former à la prise en main autonome de ces outils. Cet objectif pourra être mené en lien avec l'INIST.

- En lien avec des initiatives menées dans la communauté internationale des bibliothèques (cf Liber Data Science in Libraries Working Group), développer des outils basés sur la science des données pour le pilotage de la politique documentaire. Là aussi, un démonstrateur peut être proposé dans le cadre du programme.