Archives parlementaires de la Révolution française

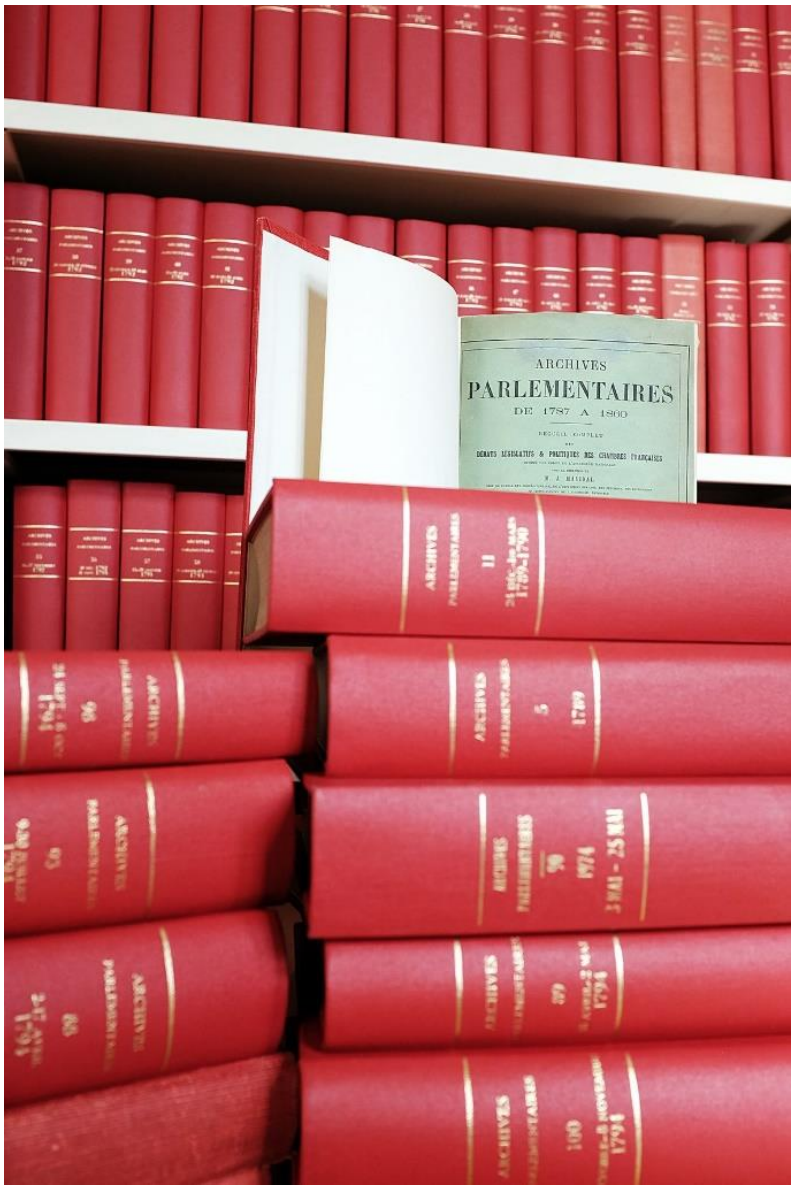
Projet lauréat de l'AAP CollEx-Persée 2018,
rapport scientifique

Hélène Bégnis, Cécile Obligi, 22/06/2021

Table des matières

1 – En amont du projet CollEx-Persée : état des lieux des réalisations antérieures

Le projet d'offrir aux internautes, et notamment aux spécialistes de la Révolution française, un accès gratuit, pratique et enrichi à la version numérisée des **102 volumes** des Archives parlementaires de la Révolution française date de 2011. En effet, à ce moment, la Bibliothèque interuniversitaire de la Sorbonne



Photographie des volumes papier de l'IHRF, cliché Lise Hébuterne, BIS

(BIS), l'unité mixte de service[1] Persée et l'Institut d'histoire de la Révolution française (IHRF)[2] s'associent pour proposer ce qui constitue la base du projet CollEx-Persée mené entre 2019 et 2021. En 2011, existaient deux versions en

[1] Devenue en janvier 2021 une unité d'appui à la recherche.

[2] Devenu en janvier 2016 une partie intégrante de l'Institut d'histoire moderne et contemporaine (IHMC).

ligne des Archives parlementaires, celle de Gallica et celle de Stanford, comportant toutes deux des inconvénients justifiant un nouveau traitement : le corpus était incomplet dans les deux cas car seuls les volumes libres de droit (1 à 82) étaient en ligne, et leur exploitation, notamment sur Gallica, n'était pas aisée. Cette publication, démarrée sous le Second Empire et toujours en cours, regroupe les comptes rendus des séances des assemblées qui se sont succédées en France à partir du printemps 1789. Le dernier volume paru[3] s'arrête au 2 décembre 1794, sous la Convention, portant la somme disponible publiée à environ 80 000 pages.

Des chercheur·e·s de l'IHRF ont défini l'outillage à apporter, la BIS s'est chargée du traitement documentaire des volumes et Persée a assuré tout le traitement informatique, de la numérisation[4] à la mise en ligne sur son portail.

L'outillage défini a été le suivant : **indexation manuelle des noms de personnes et balisage du texte.** Seuls les noms apparaissant en gras (sur les premiers volumes) ou en capitales (sur les derniers), correspondant aux débuts de prise de parole, ont été relevés, le caractère manuel du traitement, chronophage, obligeant à exclure les occurrences des noms au fil du texte. Le balisage fin du texte a été réalisé en suivant une typologie composée de 33 catégories. Il a été décidé de distinguer d'une part ce que les députés produisent et dont ils débattent, d'autre part ce que les députés reçoivent en provenance de l'extérieur de l'enceinte de l'Assemblée. La première catégorie regroupe les types suivants : amendement, appel nominal, arrêté des comités de l'Assemblée, correspondance des envoyés en mission, déclaration, proclamation et adresse de l'Assemblée, décret, loi, demande de congé, déroulement de séance, discussion, élections et nominations aux fonctions de l'Assemblée, instruction et circulaire, motion et motion d'ordre, projet de décret, projet de loi, rapport, renvoi aux comités/commissions. La deuxième catégorie contient les types suivants : adresse, pétition ou lettre envoyée à l'Assemblée, arrêté de collectivité, délibération ou procès-verbal de collectivités, discours des députations ou de citoyens à l'Assemblée, dons patriotiques et hommages, lettre. À ces deux grandes catégories s'ajoutent des documents dits « joints » : discours et opinions non prononcés, états et comptes, extraits de journaux et tableaux de données. Une dernière catégorie regroupe trois types propres à l'Ancien régime : arrêt, correspondance et discours du roi, cahiers de doléances, discours et production des ministres. Le texte est donc intégralement lu et chaque parcelle se voit attribuer l'un des 33 qualificatifs. C'est au moment où le balisage se fait que le texte issu de la **reconnaissance optique de caractères** (OCR) est attribué à chaque segment.

Le résultat du travail a été mis en ligne sur le portail généraliste Persée[5]. Le corpus a ainsi bénéficié d'une infrastructure éprouvée et bien connue donc fréquentée des internautes, mais pâti d'une structure pas entièrement adaptée à l'outillage défini. En effet, le portail permettait l'exploitation de l'indexation des

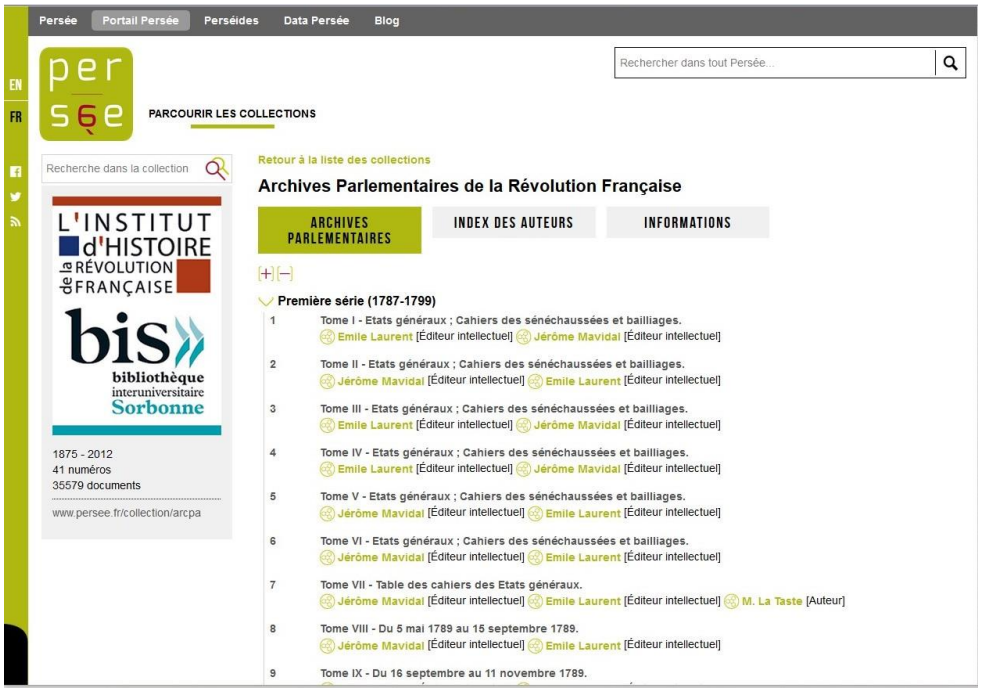
[3] *Archives parlementaires de 1787 à 1860 : recueil complet des débats législatifs et politiques des Chambres françaises. Première série, 1787 à 1799*. Tome CII, du 1er au 12 brumaire an III, 21 novembre au 2 décembre 1794 ; publié par l'Institut d'histoire de la Révolution française, Université de Paris 1 Panthéon-Sorbonne ; édité par Corinne Gomez-Le Chevanton et Françoise Brunel.

[4] Quand elle n'existait pas. Pour la majorité des volumes, Stanford a fourni gracieusement ses fichiers numériques.

[5] https://www.persee.fr/collection/arcpa.

noms, mais pas celle du balisage, dont les catégories, trop nombreuses pour l'outil, ont dû être regroupées en macro-catégories, leur faisant perdre une grande partie de leur utilité.

En 2018, 20 volumes étaient traités ; le bilan du travail faisait apparaître deux difficultés importantes : la lenteur du processus et l'inadéquation du portail généraliste Persée à un corpus spécialisé.



Archives parlementaires dans leur version en ligne sur le portail généraliste Persée, capture d'écran

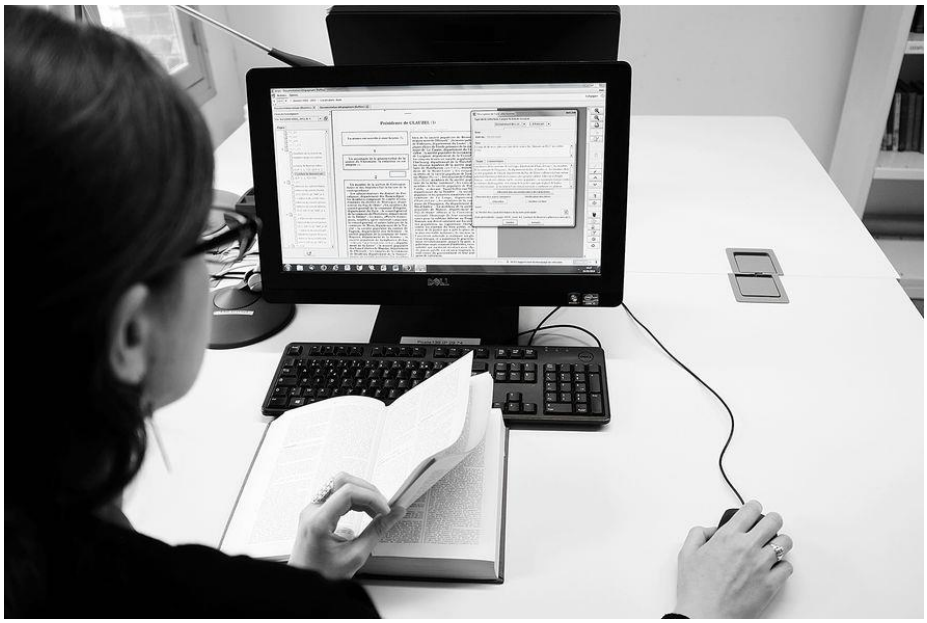
2 – Les objectifs du projet

Le projet présenté au GIS CollEx-Persée proposait la poursuite du projet par l'accélération du traitement des volumes et la mise en ligne d'un portail dédié, grâce à une subvention d'un montant total de 69 444 €. Il prévoyait trois réalisations.

Augmentation du nombre de volumes en ligne.

Au moment de la soumission du projet, le corpus en ligne comprenait 20 volumes. L'objectif était d'accélérer significativement le traitement grâce à l'embauche de 4 personnes à mi-temps pendant 10 mois, l'une des quatre personnes étant employée un peu plus longtemps (5 mois supplémentaires) pour aider à la préparation et à la gestion de la journée d'études. D'autre part, il a été décidé de commencer le traitement des volumes les plus récents (102 et antérieurs) de manière à fournir aux internautes un contenu inédit sur le web. En effet, ces volumes récents ne sont pas libres de droit donc aucun corpus en ligne

(ni Gallica, ni Stanford) n'a le droit de les mettre en ligne. L'IHMC-IHRF a accordé ces droits exclusivement pour le présent projet.



Balisage en cours dans l'application de Persée, jGalith, cliché Lise Hébuterne, BIS.

Ouverture d'un portail dédié

Le projet prévoyait la construction d'un portail dédié au corpus, que sa volumétrie et ses spécificités (balisage fin) justifiaient. Persée avait déjà développé depuis plusieurs années ce type de portails et souhaitait passer à une autre version. La perséide Archives parlementaires est la première perséide bénéficiant d'une infrastructure entièrement renouvelée.

Journée d'études

Une journée d'études, permettant de tirer de premiers résultats scientifiques du corpus, était initialement programmée le 4 mai 2020 à la Sorbonne. Du fait de la situation sanitaire, elle a été virtuelle ; d'autre part, pour permettre à Persée de présenter un dispositif de diffusion opérant, elle a dû être décalée au 31 mars 2021.

Calendrier

Initialement prévu pour une durée de 20 mois (décembre 2018-juillet 2020), le projet a dû être décalé de 8 mois et a été clôturé le 31 mars 2021, en même temps qu'a eu lieu la journée d'études.

Ce retard s'explique d'une part par la situation sanitaire, qui a engendré un retard de l'un des prestataires d'un mois. Elle s'explique surtout par la mise en œuvre, par Persée, d'une infrastructure de diffusion de la perséide entièrement repensée, développée par les équipes de Persée mais également grâce à des prestations externes spécialisées.

3 – Une nette augmentation de l'offre disponible en ligne

Le résultat a été supérieur aux attentes, à la fois parce que la productivité des agents a été forte, parce que certains volumes ont nécessité un faible temps de traitement (tables des matières) et parce qu'une partie des crédits, dédiés initialement à l'organisation en présentiel de la journée d'études, et devenant inemployables à cet effet du fait de la nécessaire virtualisation de la journée, ont été, avec l'accord du comité scientifique, utilisés pour poursuivre le traitement des volumes. Au lieu des 15 volumes supplémentaires attendus, 21 sont en ligne et 3 autres sont prêts pour intégration prochaine sur le site. L'internaute bénéficie donc en juin 2021 de 41 volumes en ligne, à comparer aux 20 disponibles avant le projet.

Sont en ligne et exploitables l'intégralité des cahiers de doléances et leur table (volumes 1 à 7), l'intégralité des séances de l'Assemblée nationale constituante et leur table (volumes 8 à 33, séances du 5 mai 1789 au 30 septembre 1791), ainsi que la toute la fin du corpus, soit les dernières séances de la Convention publiées (volumes 95 à 102, séances du 26 thermidor an II/13 août 1794 au 12 frimaire an III/2 décembre 1794).

Outre l'augmentation significative du nombre de volumes en ligne, qui permet désormais l'interrogation sur des périodes complètes (toute la Constituante), la grande amélioration pour les internautes est la mise en ligne du portail dédié, nommé perséide, qui rend possible l'exploitation de la typologie et offre des possibilités d'extraction, d'interrogation et d'études nouvelles.

Par ailleurs, un travail approfondi a été mené pour contrôler, normaliser et enrichir les données d'autorités liées aux différents intervenants ayant pris part aux séances de débats parlementaires. Pour ce faire, un chantier systématique de nettoyage des autorités (désambiguïsation de noms, corrections de dates de naissance ou de mort, normalisation des formes de noms), de curation de données (attribution à chaque intervenant d'une courte biographie) et d'alignement au référentiel IdRef (Identifiants et Référentiels pour l'enseignement supérieur et la recherche) a été réalisé. Cet alignement, automatisé et rétrospectif, conduit conjointement avec les équipes de l'Agence bibliographique de l'enseignement supérieur (ABES), a d'abord consisté à repérer les notices existantes dans data.bnf.fr (base de données sémantique contenant des données sur les œuvres, les auteurs et les thèmes du catalogue de la Bibliothèque nationale de France et de Gallica), puis à créer dans IdRef les notices manquantes à partir soit de l'exploitation des données importées directement de la BnF soit de celles fournies par Persée à partir de ses propres bases alimentées au fur et à mesure du projet et, enfin, à mettre en relation automatiquement les différents référentiels d'autorités en utilisant IdRef comme pivot. Au total, plus de 1 500 noms ont été indexés, alignés et contrôlés pour s'assurer de la pertinence et de la qualité des données. Des fiches sur chaque député mettent ainsi à disposition des internautes un ensemble d'informations documentaires agrégées telles qu'une brève biographie, une gravure, lorsqu'elle était récupérable depuis un réservoir moissonnable, et des sources externes.

4 –La perséide

Le projet a permis de concevoir et d'expérimenter non seulement une nouvelle infrastructure de diffusion, mais également de nouvelles modalités de collaborations plus inclusives par l'implication renforcée des équipes documentaires et scientifiques dans les choix mis en œuvre.

Le corpus est constitué était constitué au 31/03/2021 de 41 volumes découpés en près de 36 000 unités documentaires couvrant une période vaste allant de 1789 à 1794.

Analyse préalable et définition des besoins

En amont du développement du dispositif de diffusion, des ateliers de définition des besoins et des usages de ce corpus atypique ont été proposés par Persée à un panel d'utilisateurs·trices composé de l'ensemble des membres de l'équipe-projet de la BIS et de chercheur·es aguerri·es, doctorant·es, étudiant·es de master. L'expertise documentaire et scientifique de ce panel a été déterminante dans les choix qui ont été opérés au bénéfice de la lisibilité et de la visibilité de ce corpus.

Basés sur la méthodologie UX Design, ils ont consisté en :

- des séances de réflexion collective autour du corpus et des usages attendus ;
- la définition de cas d'usage et de profil-types d'utilisateurs·trices.

Ces ateliers ont permis de définir une architecture de l'information adaptée pour la diffusion de ce corpus sur la perséide et d'identifier les améliorations à apporter à ses modalités de consultation déjà effectives sur le portail Persée.

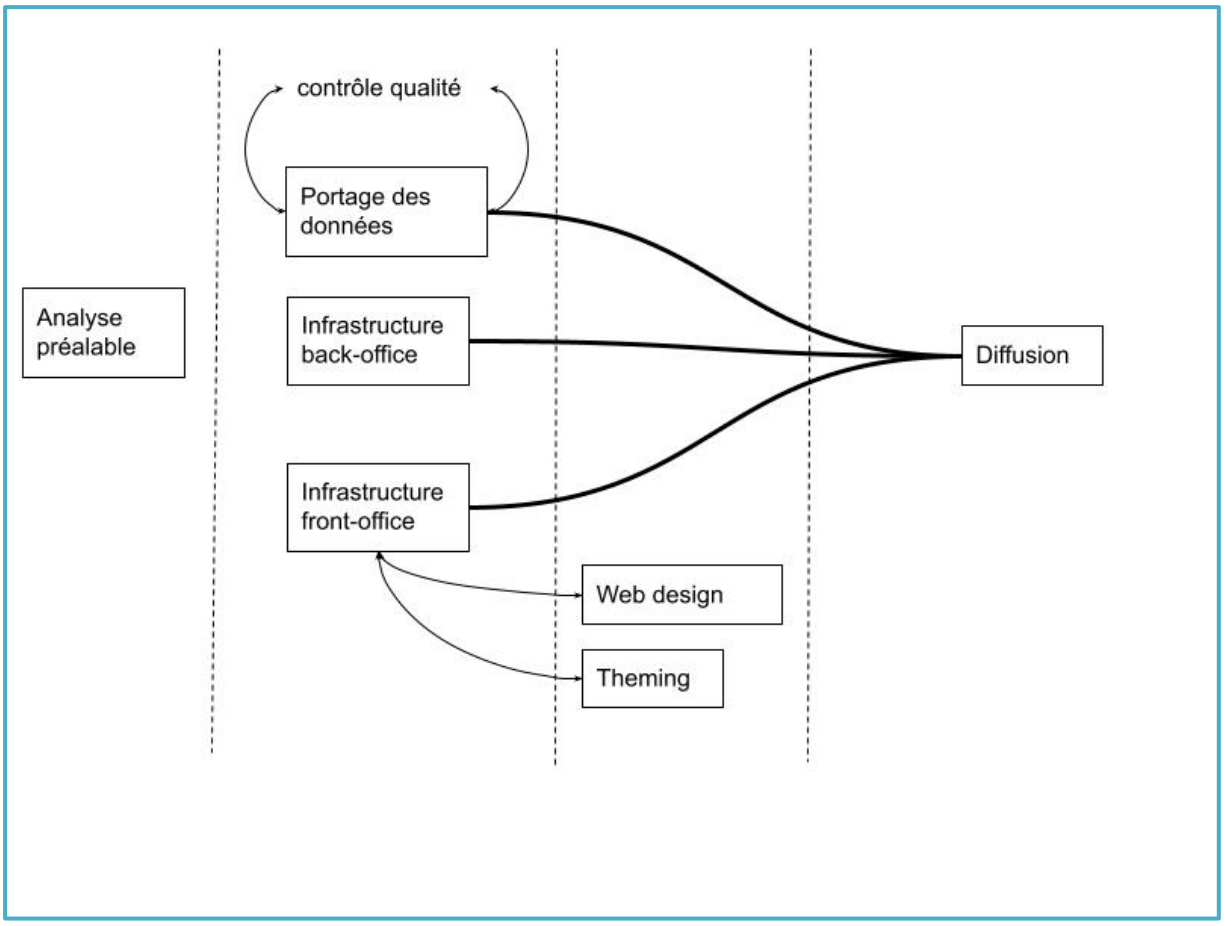
À l'issue de cette étape, les équipes de Persée ont mené une phase de définition fine des fonctionnalités attendues qui a servi de support aux développements ensuite réalisés. D'autre part, les réflexions issues de ces ateliers ont également permis de préparer la prestation de web design en travaillant les interfaces d'un point de vue fonctionnel.

Processus de développement

La démarche mise en œuvre pour le développement de cette nouvelle perséide a été la suivante :

1. Analyse préalable : étude fonctionnelle et ergonomique
2. Données : conception des scripts de portage des données produites par la chaîne de traitement de Persée pour les rendre conformes au protocole IIIF, contrôle qualité de ces données
3. Infrastructure : prise en main de Drupal, développement des scripts d'import des données produites par la chaîne de traitement de Persée et des processus de synchronisation des données et du code, mise en place de l'hébergement, définition des types de contenus Drupal
4. Interfaces : développement des fonctionnalités d'exploration du corpus (recherche, feuilletage, index, fiche document et fiche auteur), des

fonctionnalités éditoriales (contenu éditorial standard, FAQ, actualités, page d'accueil) et du système de navigation
5. Contenu éditorial : rédaction collaborative du contenu éditorial
6. Ergonomie : externalisation du web design confié à une agence web expérimentée dans la prise en charge de projets documentaires
7. Intégration : préparation des données et du code en vue de la création du thème Drupal issu de la prestation de web design
8. Diffusion :
9.

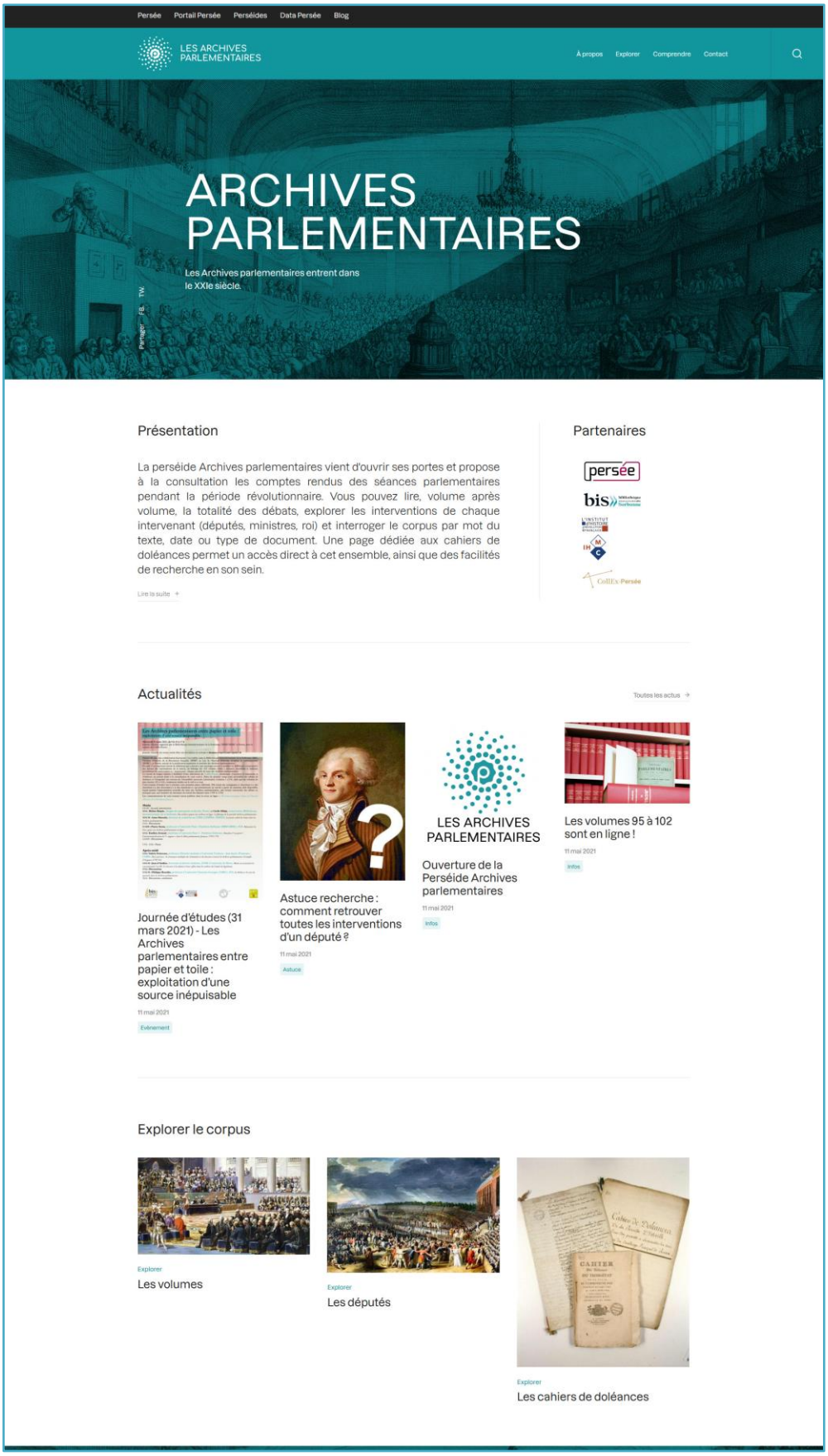


Processus de développement

Récapitulatif des fonctionnalités avant/après

Fonctionnalité	Portail Persée	Perséide
Moteur de recherche	Commun à l'ensemble des collections diffusées sur le portail Persée	Spécifique au corpus
	Typologie documentaire disponible sous forme de regroupement	Prise en charge de la totalité de la typologie documentaire
	Indexation plein-texte + métadonnées	• Indexation plein-texte et métadonnées • Recherche sur les autorités personne et les documents • Auto-complétion

		• Facettes : type de ressource, typologie documentaire, date de la séance parlementaire, volume, intervenants aux séances parlementaires
	Technologie : Sol/R + apache	• Technologie : API Search de Drupal
Consultation d'un document	Unité documentaire par unité documentaire Mode image et mode texte océrisé	• Feuilletage de l'ensemble d'un volume de la première à la dernière page • Consultation des images numérisées avec la visionneuse Mirador • Navigation facilitée par la mise en place d'une table des matières • Possibilité de se focaliser sur une seule unité documentaire à la fois • Technologie : IIIF + Mirador
Index des députés	Liste alphabétique	Liste alphabétique des députés et autres contributeurs aux séances parlementaires Affichage grille et liste Liste filtrable sur le nom du député et ses dates de vie et de mort
Liste des cahiers de doléances	Pas de liste	• Liste alphabétique des cahiers de doléances par nom de bailliage et de sénéchaussées • Affichage grille et liste • Liste filtrable sur le toponyme et le mot associé au toponyme (clergé, tiers-état, bailliage, …)
Feuilletage des volumes	Liste à plat des volumes	• Liste classée par numéro de tome des volumes • Affichage grille et liste • Liste filtrable sur les informations présentes dans le titre du volume (tomaison, date)
Fiche député	Page autorité issue de la curation de données externes obtenues grâce à l'alignement vers IdRef et des références bibliographiques des documents produits par Persée	Page autorité issue de la curation de données externes obtenues grâce à l'alignement vers IdRef et des références bibliographiques des documents produits par Persée
Apparat éditorial	Une page d'information relative au projet	Ensemble de contenus éditoriaux de description du corpus et de présentation du projet (actualités, FAQ, équipe, partenaires, …)



Page d'accueil de la perséide

5 – La communication et la valorisation du projet

La **page dédiée** au projet sur le site CollEx-Persée, https://www.collexpersee.eu/projet/archives-parlementaires-de-la-revolution-francaise/, a été alimentée et des communications régulières ont eu lieu sur les réseaux sociaux de la BIS (facebook et twitter) et via sa lettre d'actualité, notamment à l'occasion des mises en ligne de volumes.

Le projet a été présenté aux **journées professionnelles CollEx-Persée** organisées les 4 et 5 avril 2019[6].

L'**IHMC-IHRF** consacre chaque année depuis le début du projet (avant même son subventionnement par CollEx-Persée) l'une de ses séances de **séminaire** à la présentation des Archives parlementaires en ligne. En 2019, la formule a été reprise et enrichie : deux séances ont eu lieu les 20 et 27 mars 2019, elles ont permis, outre la présentation du projet, un temps de travaux pratiques et d'expérimentations qui ont donné lieu à de fructueuses remarques et suggestions d'améliorations, mises en œuvre sur la perséide pour certaines, consignées pour une évolution ultérieure pour d'autres. Depuis 2019, tout·e étudiant·e soutenant un master 1 ou 2 à l'IHMC-IHRF est tenu·e de produire une annexe à son travail de recherche justifiant des recherches effectuées sur le corpus en ligne des Archives parlementaires. Le corpus est donc de plus en plus et de mieux en mieux utilisé par son public cible.

La **journée d'études**[7], finalement virtuelle et repoussée au 31 mars, a eu lieu dans la foulée de la mise en ligne de la première version du site. Réunissant la plupart des membres du comité scientifique, dont certain·e·s ont proposé une communication, elle a permis d'une part de présenter le projet et notamment la perséide, et d'autre part d'écouter des interventions témoignant des multiples et passionnants usages possibles du corpus. Le public, atteignant un peu plus de 50 personnes, a manifesté son intérêt et formulé des remerciements pour le travail effectué. Il était principalement composé d'étudiant·e·s et de chercheur·e·s en histoire de la Révolution française.

Les sept communications de cette journée vont donner lieu à une publication dans la revue en ligne d'IHMC-IHRF, *La Révolution française. Cahiers de l'Institut d'histoire de la Révolution française*[8], dans son numéro de septembre-octobre 2021.

[6] Présentation disponible ici : https://www.collexpersee.eu/wp-content/uploads/2018/03/Pr%C3%A9sentationCollExPro19Projet2.pdf.

[7] Programme disponible ici : https://www.bis-sorbonne.fr/biu/IMG/pdf/programme_archives_parlementaires_20210331.pdf.

[8] En ligne à l'adresse suivante : https://journals.openedition.org/lrf/.

6 – Les suites du projet

Avec 41 (bientôt 44) volumes en ligne, le projet n'est pas encore terminé. Deux défis sont à relever : accélérer le traitement des volumes pour obtenir dans un bref délai l'intégralité du corpus en ligne et améliorer l'offre existante.

Garantir la complétude du corpus par l'automatisation ?

Le temps de traitement et de diffusion des volumes a connu une nette accélération : 20 premiers volumes en 8 ans, 21 volumes suivants en un peu plus d'un an. Cependant, afin d'envisager d'optimiser le temps de mise en ligne des volumes restants tout en garantissant leur exploitabilité scientifique en cohérence avec les volumes déjà diffusés, les partenaires ont engagé une réflexion visant à analyser la démarche en cours et à considérer de nouvelles modalités de traitement. L'**automatisation de l'indexation et du balisage** sont une piste en cours d'exploration. Si la recherche d'entités nommées pour les noms de personnes semble envisageable, l'automatisation du balisage respectant la typologie établie (les 33 catégories) apparaît comme plus complexe dans la mesure où elle est établie après lecture et compréhension du contenu du texte. La recherche automatisée d'entités nommées permettrait d'une part d'améliorer l'indexation en ne recherchant plus seulement les noms de personnes en gras ou en capitales (ce qui a été fait sur les 41 volumes traités), mais aussi sur les occurrences des noms à l'intérieur du texte, et d'autre part d'envisager la recherche d'autres entités nommées, géographiques, elles aussi en lien avec des référentiels nationaux.

Cette automatisation demande de traiter le corpus produit avec des outils spécifiques et de le réintégrer sur la plateforme de Persée. Une prestation extérieure est recherchée. Il importera de garantir, en amont de la prestation, la compatibilité des données produites avec la chaîne de traitement de Persée.

Améliorer le service rendu aux internautes

Le comité scientifique a relevé un problème considéré comme gênant : la **qualité de l'OCR** sur les premiers volumes. En effet, dans les volumes 1 à 33, certaines suites de caractères ont été mal reconnues et sont de ce fait peu exploitables pour une recherche plein texte. Les volumes récents en ligne (95 à 102) ne posent pas ce problème. Il importera donc de faire passer un nouvel OCR, avec un taux de reconnaissance supérieure, sans perdre le balisage effectué et en garantissant la possibilité de sa réinjection dans la chaîne de traitement de Persée.

Le comité scientifique a souligné l'importance pour lui de pouvoir effectuer des recherches et des filtrages de résultats par **noms géographiques** dans le corpus, au-delà de la recherche plein texte. Il s'agirait de s'appuyer sur un référentiel validé pour constituer un thésaurus des noms géographiques. Le travail éditorial minutieux fait pour la version papier des Archives parlementaires a permis l'identification précise de lieux dont l'exploitation serait très utile aux chercheur·e·s de la discipline. Une visualisation cartographique de certaines requêtes serait ainsi permise.

La possibilité d'**interaction** a également été demandée : pouvoir annoter le corpus, signaler des erreurs de l'édition papier (les premiers volumes papier en contiennent en effet) seraient des évolutions appréciées.

Enfin, la possibilité de **manipuler les données** en croisant plusieurs critères est apparue : par exemple croiser députés et origine géographique, sélectionner un groupe de députés et faire une recherche sur un mot/thème donné uniquement dans leurs interventions, faire des recherches à l'intérieur des interventions d'un même député. La variété des formats d'export déjà proposé par la perséide permet déjà une partie de ses actions, mais peu de chercheur·e·s sont en mesure de l'utiliser, une interface le facilitant serait une évolution intéressante.