Des archives papier aux archives en ligne : la fabrique de la perséide Archives parlementaires

Hélène Bégnis et Cécile Obligi

Édition électronique

URL : <https://journals.openedition.org/lrf/5544>

DOI : 10.4000/lrf.5544

ISSN : 2105-2557

Éditeur

IHMC - Institut d'histoire moderne et contemporaine (UMR 8066)

Référence électronique

Hélène Bégnis et Cécile Obligi, « Des archives papier aux archives en ligne : la fabrique de la perséide Archives parlementaires », *La Révolution française* [En ligne], 21 | 2021, mis en ligne le 25 octobre 2021, consulté le 27 octobre 2021. URL : <http://journals.openedition.org/lrf/5544> ; DOI : <https://doi.org/10.4000/lrf.5544>

Ce document a été généré automatiquement le 27 octobre 2021.

© La Révolution française

Des archives papier aux archives en ligne : la fabrique de la perséide Archives parlementaires

Hélène Bégnis et Cécile Obligi

Illustration 1 – Photographie des volumes papier de la bibliothèque de l'HRF, cliché de Lise Hébuterne, BIS

- 1 Depuis mars 2021, un portail donnant accès aux Archives parlementaires de la Révolution française est ouvert à l'adresse suivante : <https://archives-parlementaires.persee.fr/> Cette mise en ligne est l'aboutissement d'une longue collaboration entre la Bibliothèque interuniversitaire de la Sorbonne, Persée et l'Institut d'histoire de la Révolution française.

La genèse du projet

- 2 En 2011, la Bibliothèque interuniversitaire de la Sorbonne (BIS), l'unité mixte de service¹ Persée (ENS de Lyon et CNRS) et l'Institut d'histoire de la Révolution française (IHMC-IHRF, université Paris 1 Panthéon-Sorbonne²) se sont associés dans le but de fournir une édition gratuite, aisément accessible en ligne et augmentée de fonctionnalités dépassant la lecture linéaire ou la recherche plein texte, du vaste corpus que constituent les Archives parlementaires de la Révolution française. L'idée était de donner accès sur le web aux cent deux volumes publiés, y compris les tomes les plus récents, encore sous droits (83 à 102), en y adjoignant un outillage spécifique. Les chercheur·ses de l'IHMC-IHRF se sont chargé·es de définir cet outillage en collaboration avec la BIS, cette dernière assurant les opérations de traitement documentaire du corpus (détaillées *infra*) et Persée procédant aux traitements informatiques, de la numérisation – quand elle n'existait pas – jusqu'à la mise en ligne des volumes. De fait, le projet a pu gracieusement bénéficier de la numérisation des volumes 1 à 82, réalisée par l'université de Stanford pour son propre projet³.
- 3 Après numérisation, les volumes ont bénéficié d'une reconnaissance optique de caractères (OCR, *optical character recognition*) permettant une recherche par mots (en plein texte) et ont fait l'objet d'un traitement ayant deux visées : indexer les noms de personnes et séquencer l'intégralité du texte en unités intellectuelles cohérentes, qualifiées selon une typologie préétablie. Ces deux opérations, entièrement manuelles, ont été effectuées dans l'application que Persée utilise pour toute sa chaîne de production, jGalith (illustration 2). En raison du temps de traitement engendré par le caractère manuel du travail, l'indexation des noms s'est limitée aux noms apparaissant en gras dans les premiers volumes et en capitales dans les derniers, c'est-à-dire principalement ceux des députés prenant la parole lors des séances. Ces noms, aussi appelés « autorités personnes », alimentent le réservoir d'autorités de Persée et permettent la constitution d'un index des intervenant·es. Afin d'identifier sans ambiguïté ces « autorités personnes », Persée a fait le choix de les aligner sur le référentiel IdRef⁴, tout comme le reste de la production scientifique prise en charge par Persée. Par le jeu des rebonds que permet cet alignement, les autorités ainsi définies ont également été liées à d'autres grands référentiels nationaux et internationaux : data.bnf.fr, viaf et isni⁵. Cette opération, réalisée par Persée, était alors manuelle et se faisait au fil du traitement.

Illustration 2 – Travail en cours dans l'application de production de Persée, jGalith

- 4 Le séquençage du texte s'est appuyé quant à lui sur une typologie, définie par les chercheur·se·s de l'IHMC-IHRE, qui regroupe une trentaine de catégories⁶. Il a été décidé de distinguer, d'une part, ce que les députés produisent et dont ils débattent et, d'autre part, ce que les députés reçoivent en provenance de l'extérieur de l'enceinte de l'Assemblée. Le premier ensemble regroupe les types suivants : amendements, appels nominaux, arrêtés des comités de l'Assemblée, correspondances des envoyés en mission, déclarations, proclamations et adresses de l'Assemblée, décrets, lois, demandes de congé, déroulements de séance, discussions, élections et nominations aux fonctions de l'Assemblée, instructions et circulaires, motions et motions d'ordre, projets de décret, projets de loi, rapports, renvois aux comités/commissions. Le deuxième ensemble contient les types suivants : adresses, pétitions ou lettres envoyées à l'Assemblée, arrêtés de collectivité, délibérations ou procès-verbaux de collectivités, discours des députations ou de citoyens à l'Assemblée, dons patriotiques et hommages, lettres. À ces deux grands ensembles s'ajoutent des documents dits « joints » : discours et opinions non prononcés, états et comptes, extraits de journaux et tableaux de données. Un dernier ensemble regroupe trois types propres à l'Ancien Régime : arrêts, correspondances et discours du roi, cahiers de doléances, discours et productions des ministres. Le glossaire, disponible en ligne, propose des définitions courtes de chacune de ces trente-trois catégories. Les définitions ont trois origines : elles sont tirées des dictionnaires contemporains de l'époque, fournies par les historien·ne·s de la période, ou proposées par les concepteur·trice·s du projet, au plus près des termes et usages de l'époque révolutionnaire. Les usages de certains mots pouvant changer d'une législature à l'autre, le glossaire ne peut prétendre à l'exhaustivité ni rendre compte du caractère polymorphe de tous les termes. Quand il y a hésitation pour classer une séquence textuelle, deux principes ont prévalu : sa finalité (par exemple, un projet de décret présenté en tant que tel mais qui, adopté en l'état, est finalement typé décret) et son contenu – une allusion ou une évocation à un type de document ne suffisent pas à typer l'extrait dans la catégorie correspondante (par exemple, un document ne sera typé lettre que s'il s'agit du texte de la lettre en question ; s'il y est juste fait allusion ou qu'un résumé en est fait, il s'agira alors d'une discussion).

Illustration 3 – La collection des Archives parlementaires sur le portail Persée (<https://www.persee.fr/collection/arcpa>)

- 5 Ce travail a abouti, en 2016, à une première mise en ligne sur le portail Persée (voir illustration 3). En 2018, les tomes 1 à 20 étaient en ligne et un bilan d'étape permettait de pointer deux difficultés. D'une part, le caractère manuel du processus de balisage et d'indexation, entièrement assuré par la BIS sur ses moyens propres, rendait le travail très lent. D'autre part, la solution de mise en ligne adoptée, sur le portail généraliste Persée, si elle a fait bénéficier au projet de la notoriété et de la grande fréquentation du site, nuisait pour partie à sa visibilité mais aussi à son exploitation. La structuration du portail rend en effet impossible l'exploitation des trente-trois catégories et aucun filtre ni interrogation n'est possible par ce biais.

2018 -2021 : une nouvelle dynamique

- 6 La parution d'un appel à projet « Numérisation enrichie et services à la recherche », lancé par le groupement d'intérêt scientifique (GIS) CollEx-Persée⁷ en 2018, a permis d'inscrire l'édition des Archives parlementaires dans une nouvelle dynamique. Le projet porté par la BIS poursuivait quatre objectifs : l'accélération du traitement documentaire (indexation et balisage) par l'embauche de quatre personnes à mi-temps pour traiter rapidement une quinzaine de volumes supplémentaires ; le traitement, à rebours de l'ordre chronologique, de volumes sous droits (tome 102 et antérieurs) ; la construction d'un portail dédié, nommé « perséide »⁸, permettant de développer des fonctionnalités et modalités d'interrogation propres au corpus ; et l'organisation d'une journée d'études sur le corpus, dont les actes sont publiés dans ce numéro. Le projet a convaincu le jury, qui lui a octroyé une subvention de 70 000 € pour la période 2019-2020⁹.
- 7 Les résultats obtenus en mars 2021 dépassent certains des objectifs fixés¹⁰ : vingt-et-un volumes ont été traités, au lieu des quinze escomptés, et un travail de nettoyage

d'autorités (désambiguïsation de noms, corrections de dates de naissance ou de mort, orthographe) et d'alignement a été mené. L'alignement, cette fois automatisé et rétrospectif, a été mené conjointement avec les équipes de l'ABES¹¹. En mars 2021, plus de mille cinq cents noms ont été indexés, alignés et contrôlés pour s'assurer de la pertinence et de la qualité des données ainsi récupérées. Des fiches sur chaque député mettent ainsi à disposition des internautes un ensemble d'informations documentaires agrégées depuis des sources externes au site, telles qu'une brève biographie et une gravure – lorsqu'elle était récupérable depuis un réservoir moissonnable¹². La perséide a été mise en ligne grâce à l'investissement de Persée, qui a entièrement développé la plateforme de diffusion (basée sur Drupal pour la diffusion des contenus documentaires et éditoriaux, et sur le protocole IIIF¹³ et la visionneuse Mirador¹⁴ pour la diffusion interopérable des documents numérisés et de leurs métadonnées), tout en s'appuyant sur les travaux¹⁵ du comité scientifique réuni pour le projet¹⁶. Les interfaces utilisateurs ont été conçues par une agence web de création ayant une expertise dans la diffusion de contenus documentaires et patrimoniaux¹⁷. En mars 2021 sont traités et mis en ligne les tomes 1 à 7, correspondant à l'ensemble des cahiers de doléances ainsi qu'à leur table d'index ; les volumes 8 à 33, correspondant à l'ensemble des séances de l'Assemblée nationale constituante du 5 mai 1789 au 30 septembre 1791, ainsi qu'à leur table d'index ; les tomes 95 à 102, correspondant aux dernières séances de la Convention nationale publiées du 26 thermidor an II (13 août 1794) au 12 frimaire an III (2 décembre 1794). Le travail se poursuit sur les volumes 91 à 94 en 2021.

- 8 Ce premier corpus en ligne, bien qu'encore incomplet, rend possible une première extraction permettant de commencer à exploiter les résultats du séquençage selon les catégories définies¹⁸.

Intitulé	Nombre total	Moyenne par volume
Illustration	3	0,07
Plan	5	0,12
Extrait de journal	7	0,17
Arrêté des comités de l'Assemblée	11	0,27
À reclasser	34	0,83
Arrêté des comités et des représentants en mission	36	0,88
Appel nominal de question	57	1,39
Arrêté de collectivité	57	1,39
Note éditoriale	78	1,9
Déclaration et adresse de l'Assemblée	85	2,07
Loi	119	2,9
Arrêt, correspondance et discours du roi	158	3,85

Délibération ou procès-verbal de collectivités	172	4,2
Correspondance des envoyés en mission	234	5,71
Projet de loi	250	6,1
Instruction	303	7,39
Tableau de données	321	7,83
Demande de congés	569	13,88
Amendement	664	16,2
Élection et nomination aux fonctions de l'Assemblée	709	17,29
Discours et production des ministres	1 021	24,9
Discours des députations ou de citoyens	1 263	30,8
Discours non prononcé	1 502	36,63
Lettre	1 655	40,37
Dons patriotiques et hommages	1 682	41,02
État et comptes	2 155	52,56
Mémoire	2 158	52,63
Motion et motion d'ordre	2 428	59,22
Projet de décret	2 504	61,07
Renvoi aux comités/commissions	4 235	103,29
Rapport	6 758	164,83
Adresse, pétition et lettre envoyées à l'Assemblée	8 589	209,49
Déroulement des séances	8 592	209,56
Décret	10 038	244,83
Discussion	11 969	291,93
Total	70 427	1 717,73

- 9 Outre l'aperçu qu'il fournit sur l'ampleur du travail effectué (plus de soixante-dix mille segments de texte isolés depuis le début du projet¹⁹), ce type d'extraction, et plus

généralement la recherche par catégorie, acquerront au fur à mesure d'autant plus de pertinence que les volumes en ligne seront plus nombreux.

Illustration 4 – Page d'accueil du portail de la perséide Archives parlementaires (<https://archives-parlementaires.persee.fr/>)

- 10 Le portail en ligne (illustration 4) permet à la fois la consultation systématique (lecture linéaire d'une séance entière, d'un volume entier), l'exploration du corpus via deux interfaces dédiées, et l'interrogation par un moteur de recherche.
- 11 La liste des volumes²⁰, dont la table des matières a été intégralement saisie, permet d'accéder très rapidement à un volume voulu, puis au passage précis souhaité. Chaque séquence est téléchargeable, au format image (photographie de la version papier originale) ou au format texte (transcription issue de l'OCR).
- 12 Les travaux préparatoires ont permis de dégager le besoin de deux interfaces prioritaires pour les chercheur·se·s, la première dédiée aux cahiers de doléances, l'autre aux intervenant·e·s.
- 13 L'interface dédiée aux cahiers de doléances²¹ donne accès aux mille trois cent cinq cahiers et permet un filtrage par mot du titre du cahier : nom de lieu ou mot apparaissant dans le nom du lieu (*Ivry* ou *colonie* par exemple), baillage, sénéchaussée, tiers-état, noblesse ou clergé. Chaque cahier peut ensuite être téléchargé isolément et dispose d'un identifiant qui le rend citable (via la fonction « copier URL ») et partageable sur les réseaux sociaux (liens vers Facebook et Twitter signalés par des pictogrammes).
- 14 L'accès par intervenant·e²² – et il s'agit dans la majorité des cas d'un député – se fait par deux moyens : liste alphabétique ou recherche par nom ou partie de nom²³. Pour chaque député, le lien vers tous les réservoirs d'autorités où l'intéressé est référencé est disponible. Le rebond vers les autorités Persée permet par exemple d'obtenir toutes les occurrences de la personne, quel que soit son rôle (auteur·trice, co-auteur·trice ou sujet) dans l'ensemble des ressources présentes sur Persée (comme les *Annales historiques de la Révolution française* et la *Revue d'Histoire Moderne & Contemporaine*, pour ne citer que deux exemples parmi plusieurs centaines de milliers de sources en ligne). Sous chaque intervenant·e répertoriée sont fournis le nombre de ses « interventions » et un accès direct à chacune d'entre elles. Une intervention correspond à une occurrence du nom en gras ou en capitales. Par conséquent, si un député est interrompu deux fois dans un discours, son nom est relevé trois fois, une fois au début, deux autres fois à chaque reprise de parole. Le nombre total d'interventions, s'il constitue une première indication précieuse, est donc à nuancer par l'étude précise des textes. Enfin, la liste des « co-intervenant·e·s » donne accès à tous les orateur·trice·s intervenu·e·s durant la ou les même·s séance·s. Elle permet de savoir avec qui un député interagit le plus.
- 15 Le moteur de recherche principal, en attendant des re-paramétrages qui permettront d'exploiter toutes les informations saisies, offre déjà de riches possibilités d'exploitation. La recherche par mot (ou portion de mot²⁴) propose une autocomplétion qui suggère un ensemble de termes correspondant au mot recherché par l'utilisateur. Deux types de contenus sont indexés pour la recherche : les Documents et les Personnes. Les champs indexés pour l'autocomplétion sont :
 - Le texte intégral et le titre pour les documents ;
 - Le nom et le prénom, la qualité et la note bibliographique pour les personnes

- 16 Un traitement spécifique est appliqué au texte intégral permettant d'optimiser l'autocomplétion : seuls les chiffres, les lettres, les virgules, les espaces et les tirets sont employés pour l'indexation du texte intégral. Les majuscules sont transformées en minuscules et les lettres accentuées perdent leurs accents, et il en va de même des autres signes diacritiques. En termes d'affichage des résultats, la pertinence a été privilégiée mais le tri par titre est également proposé. Les résultats de la recherche peuvent être filtrés par type de résultat (fiche sur l'intervenant·e ou séquence du texte), par numéro de volume, par date de la séance parlementaire ou par catégorie (l'une des trente-trois définies dans la typologie). Pour chaque séquence identifiée, la consultation proposée (utilisation de la table des matières) et les fonctionnalités associées (visualisation de l'image et/ou du texte, téléchargement de l'image au format PDF et/ou du texte intégral issu de l'océrisation, citation, partage) sont ensuite les mêmes que pour les autres types de recherche.

Des perspectives d'évolution enthousiasmantes

- 17 Au terme de la phase du projet subventionnée par CollEx-Persée, plusieurs perspectives et souhaits d'évolution se dessinent. Il est nécessaire de poursuivre le travail sur les volumes restant (environ 60 % du corpus), en accélérant la vitesse de traitement. Cette accélération sera probablement à chercher du côté d'une solution au moins en partie automatisée, par recherche d'entités nommées²⁵ dans un corpus préalablement océrisé. Ces nouvelles méthodes pourraient permettre l'ajout de potentialités à ce jour non réalisables, mais fortement souhaitées par les communautés d'utilisateur·trices, notamment la recherche et la visualisation par toponyme. Enfin, la qualité de l'OCR devra être améliorée, notamment sur les volumes dont la moins bonne qualité d'impression – en particulier les premiers – a occasionné une reconnaissance des caractères moins fiable.

NOTES

1. L'UMS est devenue, en janvier 2021, une UAR, unité d'appui et de recherche.
2. L'IHRF était alors une équipe d'accueil de l'université. Depuis janvier 2016, il a intégré l'Institut d'histoire moderne et contemporaine, IHMC (UMR 8066, CNRS, ENS-PSL, université Paris 1 Panthéon-Sorbonne).
3. Projet *French Revolution Digital Archive*, remodelé depuis et dont le résultat est visible ici : <https://frda.stanford.edu/>
4. <https://www.idref.fr/>
5. <https://data.bnf.fr>, <http://viaf.org>, <https://isni.org>
6. <https://archives-parlementaires.persee.fr/fr/orientations-scientifiques>
7. <https://www.collexpersee.eu/>
8. <http://info.persee.fr/perseides/>

9. Le projet a été décalé de quelques mois et a finalement abouti en mars 2021 : <https://www.collexpersee.eu/projet/archives-parlementaires-de-la-revolution-francaise/>
10. La liste des vingt-cinq agent·te·s de la BIS, des treize agent·e·s de Persée et des sept chercheur·se·s ayant contribué au projet se trouve sur cette page : <https://archives-parlementaires.persee.fr/fr/liste-des-contributeurs-trices>
11. Agence bibliographique de l'enseignement supérieur : <https://abes.fr/>
12. Gravures récupérées depuis data.bnf.fr, [dbpedia](http://dbpedia.org) ou le site web du Sénat.
13. « International Image Interoperability Framework », cadre d'interopérabilité pour la diffusion et l'échange d'images haute résolution sur le web.
14. <https://projectmirador.org> : le code source de cet outil est libre.
15. Lesquels se sont entre autres appuyés sur des intégrant des ateliers d'*UX design*.
16. Le comité scientifique était composé de Michel Biard, Philippe Bourdin, Dominique Godineau, Hervé Leuwers, Virginie Martin et Valérie Sottocasa, sous la direction de Pierre Serna, co-pilote scientifique du projet CollEx. Émilien Arnaud et Jean d'Andlau ont également activement participé aux travaux préparatoires.
17. Danka, <https://www.danka.fr/>
18. Les cahiers de doléances sont exclus du tableau qui suit.
19. Extraction faite à la date du 26 mars 2021.
20. <https://archives-parlementaires.persee.fr/fr/explorer/les-volumes/grille>
21. <https://archives-parlementaires.persee.fr/fr/explorer/les-cahiers-de-doleances/grille>
22. <https://archives-parlementaires.persee.fr/fr/explorer/les-deputes>
23. La troncature est automatique de part et d'autre de la suite de caractères recherchée : *bois* donne accès à Dubois, Du Bois, Herbois, Boissy, etc.
24. Ici aussi, la troncature est automatique de part et d'autre de la suite de caractères recherchée : *urgen* permet d'accéder aux occurrences des déclinaisons du mot urgence (urgences, urgent, urgents, urgente, urgentes), mais aussi à celles de résurgence, par exemple.
25. Les « entités nommées » qui font l'objet de recherches sont principalement des noms de personnes, des noms géographiques ou des noms d'organisation. La recherche, faite par une machine, est connectée à un lexique prédéfini et peut porter sur un corpus écrit en langage naturel (ce qui est le cas des Archives parlementaires).

RÉSUMÉS

Un portail web permettant la consultation et l'étude des Archives parlementaires de la Révolution française vient d'ouvrir, <https://archives-parlementaires.persee.fr/> L'article contextualise le projet à l'origine de la mise en ligne, en explique les enjeux, en détaille les réalisations actuelles et en esquisse les perspectives d'évolutions futures.

A web portal has been recently opened in which it is possible to access and study the Parliamentary Archives of the French Revolution: <https://archives-parlementaires.persee.fr/> The aim of this article is to give the context of this project, from its inception to its online publication, explaining its stakes, presenting its current achievements, and sketching its prospects of improvement.

INDEX

Mots-clés : Sources, Archives parlementaires, Numérisation

Keywords : Sources, Parliamentary Papers, Digitisation

AUTEURS

HÉLÈNE BÉGNIS

Persée

CÉCILE OBLIGI

Bibliothèque interuniversitaire de la Sorbonne