SciencesPo

CollEx-Persée

Archelec4 : Exploitation du corpus d'archives électorales de Sciences Po

Projet CollEx-Persée : Archelec4
Sources primaires d'archives électorales françaises de la Vème République du Centre de recherches politiques (CEVIPOF) de Sciences Po : focus sur les élections législatives, 1958-1993.
Calendrier : Décembre 2018-juin 2021
Porteurs : Direction des ressources et de l'information scientifique (DRIS) et CEVIPOF à Sciences Po (liste complète des participants ici)

Contexte du projet

Ce projet vise à permettre aux chercheurs et aux communautés intéressées d'accéder à de nouvelles formes d'exploitation des métadonnées et à des documents à la pièce du corpus de sources primaires des archives électorales du CEVIPOF (Archelec), numérisé par la Bibliothèque de Sciences Po.

Le corpus compte 40 000 documents de nature très diverse (professions de foi, bulletins de vote, tracts, etc.), couvrant les élections françaises, présidentielle et législatives, de 1958 à 2012, disponibles en accès ouvert sur la plateforme Internet Archive. Le projet Archelec4 se focalise sur les élections législatives de 1958 à 1993, soit 31 943 professions de foi concernant 63 886 candidats sur 9 élections.

La qualité des métadonnées proposées est un enjeu essentiel du projet pour pouvoir envisager ensuite de les croiser ou de les lier à d'autres corpus, ou pouvoir comparer plusieurs sources de données, par exemple le bord politique issu du déclaratif des professions de foi ou transmis par le ministère de l'Intérieur, provenant des livres blancs ou de la base EDEN du CEVIPOF. La quantité de métadonnées permet d'enrichir la qualité des travaux de recherche, notamment sur les spécificités des candidats (et suppléants) : sexe, profession et parti ou autre organisation politique, que ce soit sur la collection des professions de foi ou sur la collection des logos des partis, nouvelle collection issue de ce projet (Cf. infra). La liste des métadonnées est mise en annexe à ce rapport.

Archelec4 s'inscrit dans la continuité des projets de numérisation des appels à projets 2013 de la Bibliothèque scientifique numérique (BSN) BSN5 Archelec1, puis Archelec2 et Archelec3, et Archelec5 (sur les élections européennes) réalisés sur des ressources propres à Sciences Po. L'objectif d'Archelec4 est d'adapter le corpus aux nouveaux usages des humanités numériques tout en diffusant plus largement les données électorales.

Objectifs du projet

Le projet vise à permettre à la communauté de recherche :

- de bénéficier d'un accès à la pièce aux professions de foi et aux bulletins de vote, permettant aussi bien la recherche plein texte combinée avec un ou plusieurs critères que la constitution de bases de données.

- de mener des recherches sur les documents mis en ligne grâce à de nouvelles métadonnées correspondant aux besoins des chercheurs et aux informations disponibles sur les documents.

- d'apporter une nouvelle dimension d'analyse de contenu du corpus à travers une collection de logos des partis et organisations politiques pendant la période couverte.

- d'expérimenter une nouvelle forme d'exploration des contenus et d'envisager leur croisement avec d'autres sources comme les résultats électoraux.

- d'offrir un accompagnement à l'utilisation du corpus.

Les chiffres clés du projet sont proposés dans l'infographie en annexe de ce rapport.

Indexation et fouille de texte : un équilibre à trouver

La projection des modes d'exploitation du corpus implique à la fois la consultation des utilisateurs potentiels, principalement des chercheuses et des chercheurs, mais également l'expérimentation d'outils d'exploration avant de les utiliser sur le corpus.

L'équipe a donc tout à la fois entamé un dialogue avec les chercheurs membres du conseil scientifique du projet sur le choix des métadonnées à extraire manuellement des professions de foi et a sollicité également la contribution d'un expert pour expérimenter l'extraction d'entités, à partir des mêmes professions de foi, ainsi que l'émergence de clusters (partitionnements) ou de tendances à partir de ces données en volume.

Selon les chercheurs consultés, tous spécialistes de science politique, les besoins et les avis sur l'indexation diffèrent. Le cas de la métadonnée "profession" (du candidat ou du suppléant) est représentatif de la difficulté à arbitrer :

- soit s'en tenir au déclaratif, c'est-à-dire indexer la profession des candidats telle qu'elle est mentionnée, le cas échéant, sur le document ;
- soit ajouter un niveau d'indexation générique - le codage - et, éventuellement, adosser ce niveau d'indexation à un référentiel, dans le cas présent, à une nomenclature métier ;
- soit ne pas en tenir compte au niveau de l'indexation (pour pouvoir se concentrer sur d'autres indexations) et privilégier les outils de fouille de texte pour extraire des termes relatifs aux professions et métiers.

Les arbitrages ont tenu compte de ces besoins et recommandations, mais également des besoins connus des autres catégories d'utilisateurs du corpus : journalistes, archivistes, généalogistes, iconographes, citoyens ou encore enseignants. Les arbitrages ont enfin été effectués en fonction du temps dont a disposé l'équipe et des moyens de faire appel à des vacataires indexeurs.

D'autres arbitrages ont été menés de la même manière pour la collection des logos des partis politiques tels qu'ils sont présents sur les professions de foi. En anticipant les usages, un consensus s'est dégagé pour sélectionner et extraire un seul exemple d'un logo présent à de nombreuses reprises et un minimum de métadonnées utiles à ce stade.

Pour adapter la réponse aux besoins des chercheurs, il a été décidé de donner la possibilité de télécharger tout ou partie du corpus de données, c'est-à-dire les métadonnées et les fichiers océrisés, et de guider les chercheurs dans ces procédures. S'est posée ensuite la question de l'accompagnement des chercheurs à l'exploitation de ces données : avec quels outils et selon quelles modalités d'assistance ? C'est dans ce cadre que nous avons fait appel au concours de Jean-Philippe Moreux, expert scientifique de Gallica à la Bibliothèque nationale de France (BnF), qui a testé des méthodes et outils d'extraction d'entités à partir de la suite GoogleCloud (illustration 1) et de CorText et d'annotations grâce à l'outil IBM Watson "Natural Language Understanding" (illustration 2).

https://cloud.google.com/vision/

https://natural-language-understanding-demo.ng.bluemix.net/

A Front National	Organization	0.95
Assemblée nationale	Organization	0.94
Front National	Organization	0.89
R.P.R. et de l'U.D.F.	Person	0.79
R.P.R., U.D.F., P.S., P.C.	Person	0.61
7	Number	0.41
agriculteurs	JobTitle	0.37
France	Location	0.36
députés	JobTitle	0.34

Même si les résultats ne sont pas exploitables pour alimenter la collecte de métadonnées dans le cadre du projet Archelec4, une expérimentation s'avère en revanche envisageable, du type analyse en fréquence de termes choisis sur la totalité du corpus. La pandémie a empêché l'organisation d'un événement autour de ce type d'expérimentation, événement qui aurait dû être précédé par une appropriation minimum des outils par les porteurs du projet, pour pouvoir identifier et inviter des ingénieurs et des chercheurs intéressés.

Traitement des données et mémoire des projets

La problématique de la mise à disposition des professions de foi à la pièce témoigne de la complexité de la gestion du type de projet que représente Archelec4. Un de ses objectifs était de découper à la pièce les fichiers pdf des professions de foi numérisées qui étaient jusqu'alors rassemblées en recueils par tour d'élection dans une circonscription, et de leur attribuer à chacune les métadonnées correspondantes (nom du candidat, du suppléant, parti, soutien, etc.).

Ce travail de découpage a nécessité de ré-identifier de nombreux acteurs qui avaient travaillé sur les précédents projets de numérisation et d'indexation[1], de retrouver la trace des traitements appliqués aux documents à l'époque, de l'endroit où ils ont été conservés et des standards utilisés. Parmi ces acteurs pouvaient figurer des collègues mais également des prestataires extérieurs. La difficulté que l'ingénieur de l'équipe projet a éprouvé à retrouver la trace des fichiers des professions de foi numérisées (avant qu'ils ne soient rassemblés et diffusés) témoigne de cette complexité et de la nécessité de documenter un projet au-delà de son rapport scientifique. C'est une tâche à expliciter, un peu à la manière dont la gestion des données de la recherche est désormais anticipée grâce au plan de gestion des données (PGD). C'est à partir des fichiers bruts issus de la numérisation par le prestataire qui avait assuré la numérisation lors des précédents projets Archelec que les recueils de professions de foi ont finalement pu être divisés en documents à la pièce.

Parallèlement, une problématique technique relative au nommage des professions de foi s'est posée pour en assurer l'interopérabilité entre les outils et entrepôts : Internet Archive, NumaHop (cf ci-dessous)), mais aussi la bibliothèque numérique de Sciences Po et la base des résultats électoraux, à terme. A noter que NumaHop, plateforme de gestion de chaînes de numérisation mise en œuvre par la bibliothèque Sainte-Geneviève, la bibliothèque de Sciences Po et la BULAC, a permis la remise en ligne des 30 000 PF à la pièce en y associant les nouvelles métadonnées (https://www.numahop.fr/).

Un schéma synthétique sur le circuit de l'information, le transfert des documents et les interventions de chaque interlocuteur est proposé sur la page suivante.

[1] Rappelons ici qu'Archelec constitue une suite de projets depuis Archelec1 jusqu'à Archelec6 en cours actuellement.

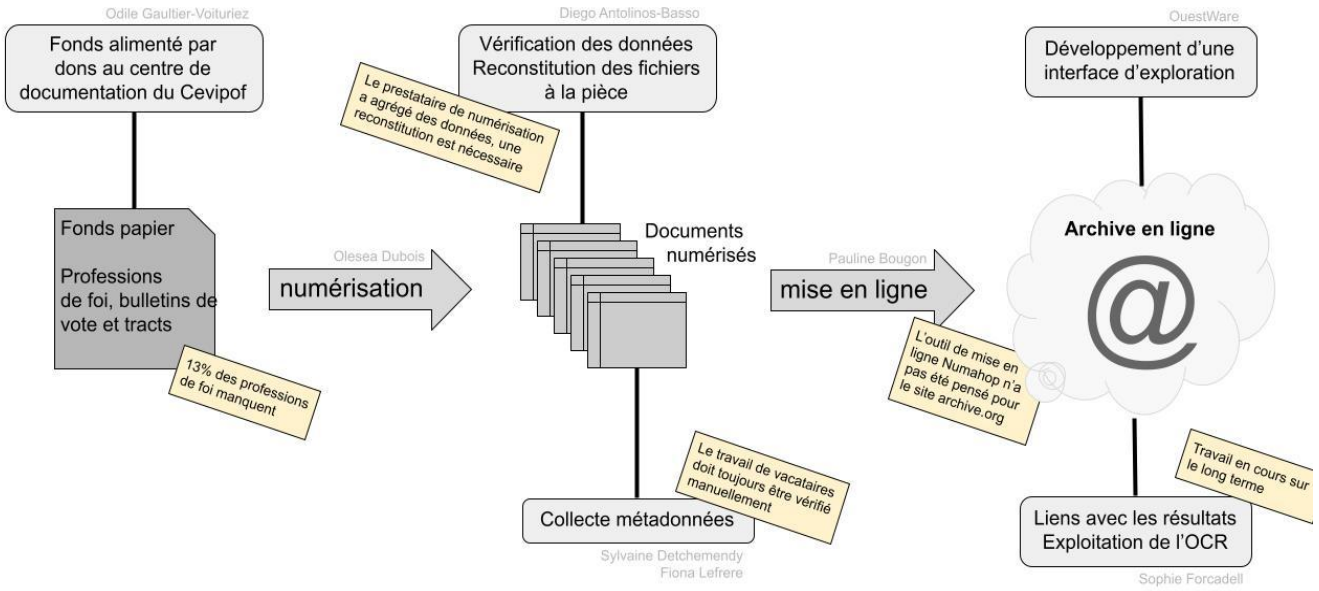


Schéma : circuit informations et transfert documents. Apport des compétences/personnes à chaque étape du projet. Flux et relations.

Approche d'une interface par prototypage

À partir du constat de l'équipe projet que l'interface d'accès au corpus d'Internet Archive ne permettait pas d'exploiter les métadonnées ajoutées au corpus et la manipulation des données en volume, et à partir de la collecte des besoins des chercheurs, il a été décidé, assez rapidement dans l'avancée du projet, de procéder à un prototype d'interface alternative d'exploration du corpus afin de donner aux utilisateurs un premier aperçu des principaux bénéfices d'un tel outil.

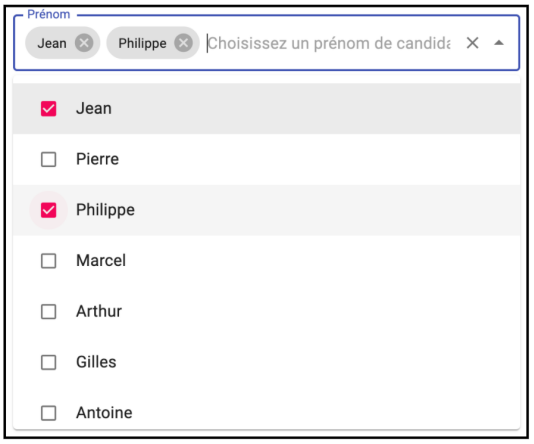
L'approche par prototype visait à obtenir rapidement une première vue concrète de ce que pourrait être une interface afin d'arbitrer sur son développement et donc sur les moyens à prévoir. C'est dans cette optique qu'un stagiaire venant de l'université de Compiègne, en 2è année, a été recruté. Son profil était moins expert en développement qu'en bonne compréhension technique de la complexité du projet, des attendus des

chercheurs et il a démontré sa capacité à anticiper les principales problématiques techniques et à produire des premières maquettes de l'interface (illustrations 3 et 4).

Maquette de visualisation :
- d'une profession de foi avec points d'entrée sur les catégories de métadonnées
- d'une liste de professions de foi

- d'une interface de recherche avancée exploitant les nouvelles métadonnées du corpus.

(page de résultats d'une recherche)



Maquette du principe d'interrogation des champs de recherche (qui intègre l'auto-complétion)

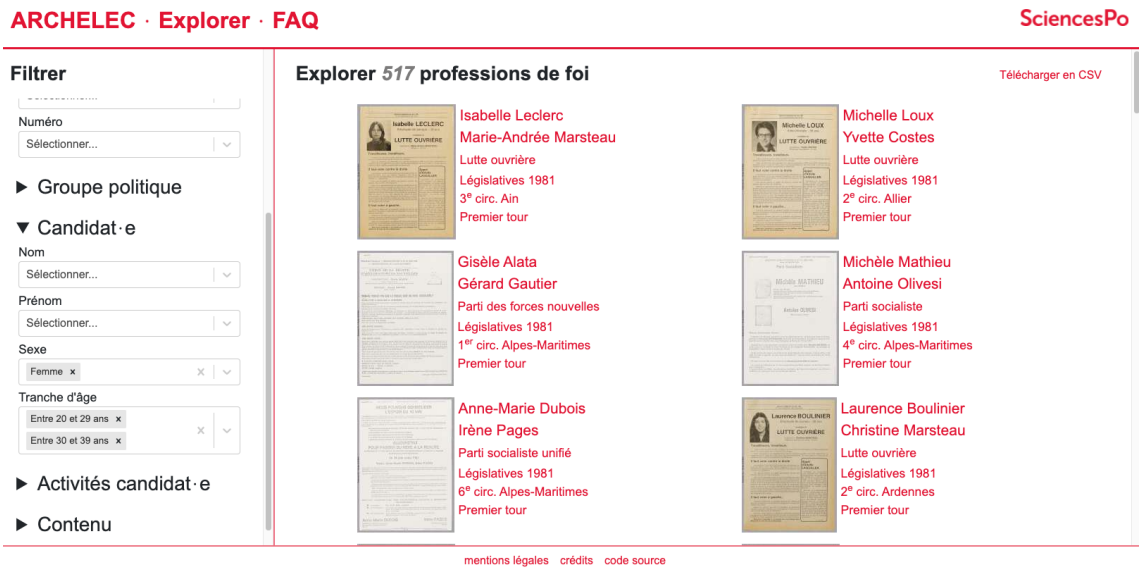
Notre stagiaire n'a pu profiter de la présence de l'équipe dans les murs qu'un petit mois avant de télétravailler pour cause de pandémie dès la mi-mars 2020. L'ambition de formation qu'avait l'équipe, et plus particulièrement l'ingénieur données, à son égard a dû être revue à la baisse et la compréhension du projet dans toute sa complexité et son historique n'a pas été simple à transmettre. En effet, au-delà d'une documentation et d'une mémoire du projet à mettre à la disposition de tout nouvel entrant (coéquipier, stagiaire, vacataire ou prestataire), la transmission orale et la coopération en mode atelier sont tout aussi importantes et ces possibilités ont manqué pendant la pandémie.

Combiner recherche avancée, exploration et visualisation

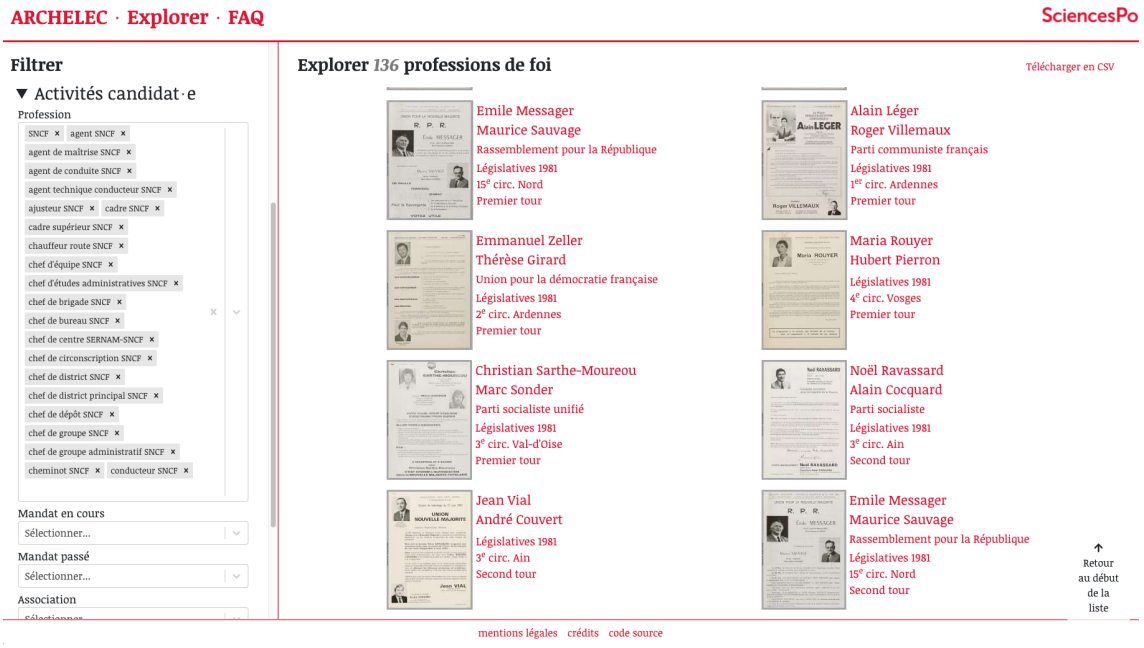
L'aboutissement des discussions et arbitrages scientifiques s'est incarné dans le cahier des charges du développement de l'interface d'exploration du corpus et de visualisation des données en volume. L'équipe projet a expérimenté à cette occasion le dialogue fécond avec le prestataire, déjà rompu à ces techniques (interface de recherche, ergonomie, design d'interface) et à plusieurs méthodes de visualisation de données (frise chronologique, matrice, etc.).

La méthode agile, par ateliers successifs, s'est avérée productive, permettant d'alimenter les développeurs en remarques, corrections et suggestions dès les premières étapes, et ce, sur une grande partie du projet. L'intervention de cet acteur tiers a par ailleurs permis de replacer le projet dans un contexte plus large et de réfléchir à la clarté des libellés utilisés sur l'interface, ainsi qu'au périmètre et aux contenus de la FAQ pour s'adresser au plus grand nombre. Là encore, trouver le bon

équilibre entre sophistication de la recherche avancée et volonté de ne pas surcharger le design de l'interface a nécessité de remettre en perspective les usages connus et anticipés de l'outil, ainsi que des pistes d'évolution futures, restant à préciser après des premiers retours d'utilisateurs (illustrations 5 et 6).



Affichage des résultats de recherche à partir de [l'interface de recherche Archelec Explorer](). Requête : Elections 1981 ou 1988 AND Candidat Femme AND Tranches d'âge Entre 20 et 29 ans ou Entre 30 et 39 ans.



Affichage des résultats de recherche à partir de l'interface de recherche Archelec Explorer. Requête : Activité candidat contient SNCF.

Notre objectif a été ainsi clarifié : donner suffisamment de possibilités de recherche pour permettre à tout utilisateur de retrouver facilement une ou plusieurs professions de foi et aux chercheurs d'explorer ou d'exporter des sous-corpus pour du traitement et de l'analyse de données en volume, et pour faire émerger de nouvelles problématiques de recherche.

Enfin, la visualisation des données en grandes quantités constitue une ou plusieurs fonctionnalités que les chercheurs pourraient utiliser directement sur l'interface, tout en sachant que, pour les plus aguerris d'entre eux, il leur est possible de déployer leurs propres méthodes et outils après téléchargement de tout ou partie du corpus. Les différents types de visualisation doivent par ailleurs constituer une manière pour les autres utilisateurs d'appréhender le corpus sous une autre forme que celle de la recherche pièce à pièce (ou lot par lot). On retrouve ici l'importance de la discussion et des arbitrages pour trouver le meilleur équilibre et intégrer le public le plus large possible.

Pour autant, le projet ayant ses limites, notamment financières, l'ambition de départ qui consistait à préférer une approche géographique et cartographique aux visualisations n'a pas pu s'inscrire dans ce projet. En effet, le paysage français des circonscriptions électorales a connu de nombreuses évolutions depuis 1958 (date des premières élections présentes dans ce corpus) et un travail très affiné de cartographie doit être mené pour pouvoir visualiser ces évolutions tout en conservant la possibilité de comparer des élections entre elles.

Conclusion : Préparer le futur

Les bénéfices du projet vont plus loin que les seuls ajouts de métadonnées au corpus, leur accès à la pièce, et la création d'une interface de recherche et de visualisation ergonomiques et adaptées aux besoins des chercheurs. En effet, la coopération avec les chercheurs et un ingénieur de recherche, ainsi qu'avec le prestataire informatique, a apporté un dialogue riche pour arbitrer sur les choix tout au long du projet mais aussi pour envisager les futurs projets Archelec.

Le dialogue nous a notamment permis d'explorer les méthodes à envisager pour pouvoir croiser les données du corpus avec les résultats électoraux obtenus par les candidats (c'est-à-dire par les professions de foi !). Ce croisement de données est en réalité complexe car il suppose de considérer le nom du candidat à la place du numéro de la profession de foi, utilisé aujourd'hui comme donnée-pivot.

Un deuxième besoin s'est fait jour, grâce à la possibilité d'exploitation des photos des candidats, qui serait d'opérer une reconnaissance automatique de formes indiquant des sentiments ou encore d'exploiter les couleurs des professions de foi et les comparer aux partis politiques.

Par ailleurs, compte tenu de l'aspect chronophage et coûteux des projets d'enrichissement du corpus[2], à la fois en nombre de documents sous forme numérique (numérisation) et en métadonnées (ajout manuel ou extraction), l'intégration de ces procédures et manipulations devraient continuer à être documentés à l'avenir et intégrés au montage de futurs projets de recherche. On pourrait ainsi envisager de les communiquer sur la plateforme Archelec Explorer, sur une page dédiée aux projets de recherche. Enfin, ce projet a mis à jour le potentiel de la contribution citoyenne à l'enrichissement du corpus, au signalement des éventuelles erreurs et au recueil des observations de tous les utilisateurs.

De futurs projets nous permettraient de prolonger le corpus de manière longitudinale et en ajoutant d'autres types d'élections, régionales ou départementales par exemple. Le partenariat avec les Archives nationales pourrait être renforcé pour compléter les professions de foi manquantes.

Ressources utiles

Archelec Explorer, interface de recherche avancée et de visualisation du corpus Archelec.

Corpus Archelec sur Internet Archive.
Corpus Archelec sur la bibliothèque numérique de Sciences Po.

Présentation du projet lors des journées professionnelles CollEx des 4 et 5 avril 2019. Vidéo de 2 minutes.

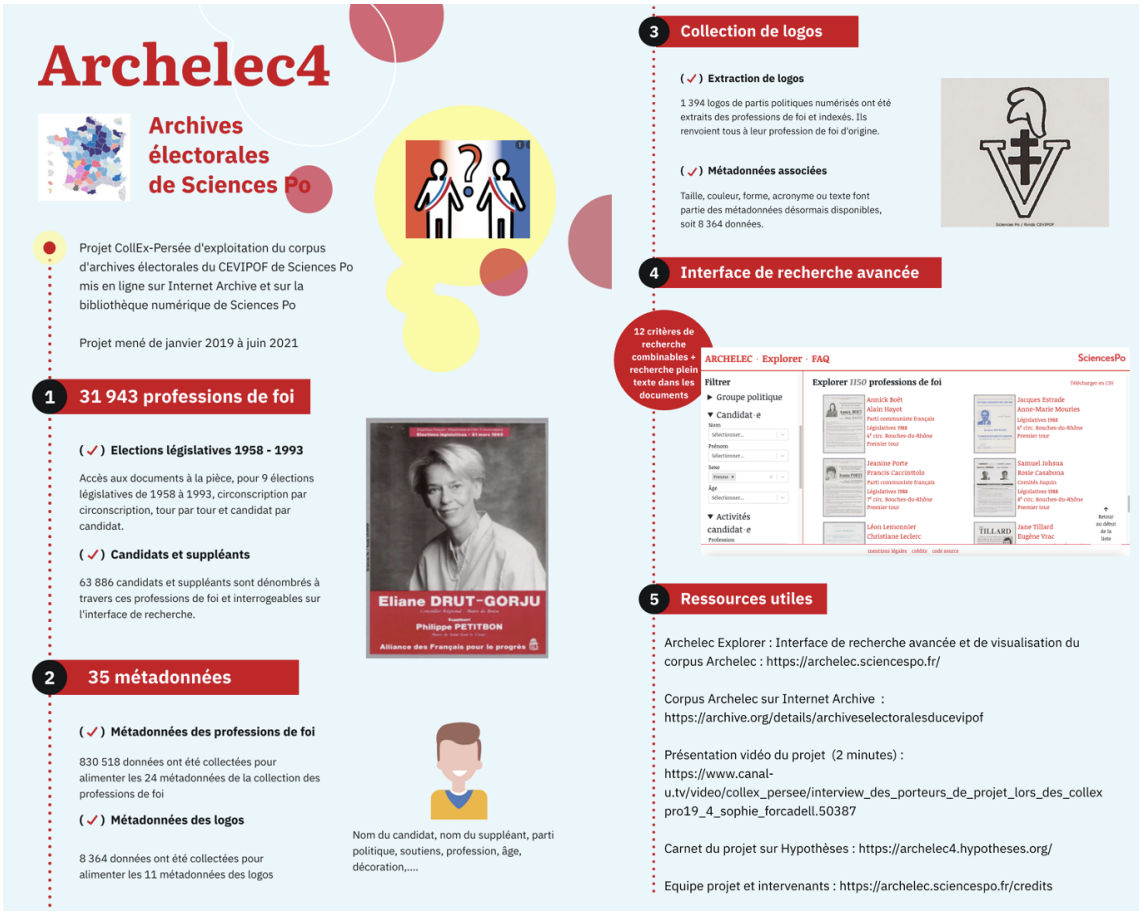
[2] Il est par exemple envisagé d'ajouter au corpus les élections présidentielles, européennes et régionales depuis 1993.

Carnet du projet sur Hypothèses.

Equipe projet et intervenants

Annexes

Chiffres clés du projet (également en pièce jointe au rapport)



Liste des métadonnées du corpus de professions de foi

Métadonnées préexistantes, issues des précédents projets Archelec	Métadonnées ajoutées lors du projet Archelec 4

N° identifiant	Rang
Code boite (d'archive)	Nom
Election	Prénom
Année	Sexe
Mois	Age / Année de naissance
Département - code	Profession
Département - nom	Mandat public en cours
Circonscription	Mandat public passé
Tour	Activité associative
Type de document	Autre statut
Numéro de document	Décorations
	Partis en soutien
	Liste

Liste des métadonnées de la collection des logos

Métadonnées créées lors du projet Archelec 4	Commentaires
dc:identifier	L'identifiant est concaténé à partir des informations "cote boîte" à "n° de document".
cote boite	
election	
annee	
mois	
departement_code	

departement_nom	
circonscription	
tour	
type_de_document	
numero_document	
organisation	L'organisation = parti est reprise et visible par l'usager dans le title (ainsi que le type d'élection et l'année).
dc/title	
impression	De cette métadonnée "impression" jusqu'à la métadonnée "texte", ce sont les données descriptives du logo.
couleur	
objet_sujet	
forme	
acronyme	
texte	
dc:relation	Renvoie au document original.
dc:description	Correspond à une "note" concernant le logo.
dc:subject	Ces 4 dernières métadonnées relèvent de la présentation en ligne sur Internet Archive.
dc:subject	
dc:subject	
dc:type	