h e g

Haute école de gestion
Genève

La couverture de l'archivage du web suisse : comparaison des approches de la Bibliothèque nationale suisse et d'Internet Archive

Rapport final

Projet de recherche réalisé par :

Christelle DONIUS

Anna HUG BUFFO

Sous la direction de :

Arnaud GAUDINAT, Professeur HES

Carouge, le 15 janvier 2020

Master en Sciences de l'information

Haute École de Gestion de Genève (HEG-GE)

Hes·so GENÈVE
Haute Ecole Spécialisée de Suisse occidentale

Déclaration

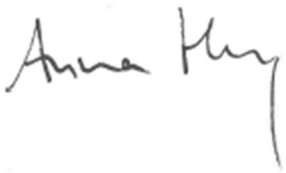
Ce travail de recherche est réalisé dans le cadre du Master en Sciences de l'information de la Haute école de gestion de Genève. Les étudiantes acceptent, le cas échéant, la clause de confidentialité. L'utilisation des conclusions et recommandations formulées dans ce travail, sans préjuger de leur valeur, n'engage ni la responsabilité des auteurs, ni celle de l'encadrant.

« Nous attestons avoir réalisé le présent travail sans avoir utilisé des sources autres que celles citées dans la bibliographie. »

Fait à Carouge, le 15 janvier 2020

Christelle Donius

Anna Hug Buffo



Remerciements

Nous remercions Arnaud Gaudinat, notre professeur encadrant, pour ses encouragements réguliers, ainsi que Bastien Berger, assistant à la HEG-ID, pour son soutien technique indispensable à la création des scripts d'interrogation pour la récupération des données.

De même, Michael Marti nous a aidées pour paramétrer les analyses des données.

Un énorme merci à Barbara Signori, Maya Bangerter et l'équipe technique de la Bibliothèque nationale, qui ont aimablement extrait les données des Archives Web Suisse à notre demande.

Finalement, nos remerciements vont à nos proches, qui nous ont apporté un soutien moral pendant toute la durée de ce projet de recherche, et parfois un brin d'inspiration insoupçonné.

Résumé

Le web est devenu indispensable dans notre société actuelle centrée autour de l'information et de la communication. La valeur patrimoniale d'au moins une partie de ses contenus est indiscutable. Mais il s'agit de supports volatiles et techniquement difficiles à traiter, et les volumes sont énormes.

Ce projet de recherche s'intéresse à la couverture de l'archivage du web suisse par deux acteurs, la Bibliothèque nationale suisse (BN) d'un côté et Internet Archive (IA) de l'autre. Du point de vue organisationnel, la différence majeure entre les deux institutions est que la BN a une approche sélective, tandis qu'IA moissonne tous les contenus rencontrés par ses *crawlers*, sans distinction qualitative. Le concept de "web suisse" englobe, pour nous, les sites correspondant à la définition des "Helvetica" utilisée par la BN.

Nous avons formulé une demande auprès de l'institution (BN) et interrogé l'API disponible à cet effet (IA) pour obtenir les données brutes nécessaires à nos recherches, à savoir des fichiers CDX et XML avec les métadonnées sur les sites moissonnés. Nous les avons travaillées et analysées à l'aide du logiciel Dataiku, pour ne conserver que les données des premières captures des domaines de premier niveau.

Ainsi, à fin 2019, sur un total de 2'259'952 sites avec le ccTLD .ch, IA en archive 1'298'225 (57.44 %) et la BN 7'513 (0.33 %). 7'418 sites sont archivés par les deux institutions. Si l'on regarde les collections de la BN tous TLD confondus, 8'132 sites sont archivés. Sur ces URL, 8'048 sites se trouvent également chez IA.

Ces analyses quantitatives ont été complétées par une exploration qualitative des contenus archivés pour un échantillon de 23 sites. Nous avons vérifié leur présence dans les deux archives du web. Sur les 23 sites examinés, 10 sont archivés par la BN et 22 par IA.

IA couvre le web suisse plus largement que la BN. Mais si un site a été sélectionné par la BN pour archivage, il sera alors archivé avec un niveau qualitatif très élevé. Nous pensons que les deux approches – sélectivo-qualitative et moissonnage massif mais moins profond – sont complémentaires et répondent aux objectifs fixés par chacune des institutions.

Mots-clefs : archives du web – archives électroniques – Bibliothèque nationale suisse – Helvetica – Internet Archive – page web – site web

Table des matières

Liste des tableaux

Liste des figures

Glossaire

API (Application Programming Interface) : “A set of routines, protocols and tools for building software applications; specifically, establishing the interface (calling conventions) by which a software application accesses the operating system and other services.” (InterPARES 2020). Les API permettent à des programmes informatiques d’interagir, par exemple pour consulter des données *via* des requêtes HTTP.

Archivage du web : “Web archiving refers to the activities of selecting, capturing, storing, preserving and managing access to snapshots of Internet resources over time.” (Organisation Internationale de Normalisation 2013).

CDX : format de fichier balisé pour les métadonnées descriptives de documents web archivés. La première ligne d’un document CDX fournit la légende, sous forme codée, des colonnes du fichier, afin d’identifier la nature des données contenues dans les lignes suivantes (IIPC 2015).

Dataiku : logiciel de traitement et d’analyse de données. Les fonctionnalités que nous avons utilisées sont la rédaction de scripts, le chargement, le nettoyage et le traitement de données, ainsi que certains outils d’analyse.

JSON (JavaScript Object Notation) : format de données textuelles permettant la représentation d’informations structurées (JavaScript Object Notation 2020).

MD5 : l’algorithme Message Digest 5 est une fonction de hachage cryptographique qui permet d’obtenir l’empreinte numérique d’un fichier (MD5 2019). Il est entre autres utilisé pour déterminer si deux fichiers informatiques sont identiques, par exemple pour vérifier l’intégrité d’un document après sa transmission électronique.

SURT (Sort-friendly URL Recording Transform) : retranscription des noms de domaines permettant de faciliter le tri et la recherche. Les éléments de l’URL séparés par un point sont énumérés en allant du général vers le particulier (donc en ordre inverse par rapport à une URL standard). Par exemple, le nom de domaine *hesge.ch* sera transcrit “ch,hesge)/” tandis que seco.admin.ch sera transcrit “ch,admin,seco)/”. En revanche, une partie d’URL qui suit une barre oblique est conservée telle quelle. Par exemple : *collectif440hz.ch/projets/blindtest/* sera transcrit “ch,collectif440hz)/projets/blindtest/” (Internet Archive 2018).

Top level domain / TLD : domaine de premier niveau en français. Le niveau plus haut dans le système de nommage des entités connectées à Internet (*domain name system* / DNS). Ils peuvent être d’ordre national (country code top level domains ou ccTLD : .ch, .fr, .fi…) ou général (generic top level domains ou gTLD : .org, .com…) (Organisation Internationale de Normalisation 2013).

WARC (Web ARChive format) : standard ISO 28500:2017. “File format that specifies a method for combining multiple digital resources into an aggregate archival file together with related information”. (Organisation Internationale de Normalisation 2013).

Liste des sigles et abréviations

API	Application Programming Interface
BN	Bibliothèque nationale suisse
ccTLD	country code Top Level Domain
FTP	File Transfer Protocol
gTLD	generic Top Level Domain
HEG	Haute école de gestion
HTML	Hypertext Markup Language
HTTP	Hypertext Transfer Protocol
IA	Internet Archive
IIPC	International Internet Preservation Consortium
ISO	International Standard Organization (Organisation Internationale de Normalisation)
JSON	JavaScript Object Notation
MD5	Message Digest 5
OFCOM	Office fédéral de la communication
SURT	Sort-friendly URL Recording Transform
TLD	Top Level Domain
URI	Uniform Resource Identifier
URL	Uniform Resource Locator
URN	Uniform Resource Name
WARC	Web ARChive format
XML	eXtended Markup Language

1. Introduction

1.1 Contexte

Depuis une vingtaine d'années, Internet est devenu un vecteur de communication prépondérant dans notre société. Le World Wide Web (ci-après "web") en particulier contient un volume d'informations colossal : plus de 5,5 milliards de pages sont indexées (WorldWideWebSize.com 2019), et le "deep web", la partie qui n'est pas accessible aux moteurs de recherche, est plus grand encore. Le contenu de ces pages reflète toutes les occupations humaines, des cours boursiers aux ragots sur les célébrités, de la législation aux recettes de cuisine, des jeux en ligne aux gestes de premier secours. Il est diffusé selon une pluralité de technologies et de formats de fichiers, notamment multimédia.

Une particularité du web est sa volatilité. Des pages sont créées ou abandonnées, les contenus sont mis à jour, les noms de domaine changent. Selon la méthode d'évaluation utilisée, la durée de vie moyenne d'une page web est estimée à 44 jours (Kahle 1997), 75 jours (Guy 2009), ou encore à 2 ans et 7 mois (Crestodina 2017) ; une étude a montré que près d'un tiers des liens cités dans les articles scientifiques n'étaient plus fonctionnels, 5 ans après publication (Sampath Kumar et al. 2015). Avec le web 2.0 et ses outils interactifs, les modifications sont devenues beaucoup plus fréquentes encore.

D'un point de vue patrimonial, la question de l'archivage de tous ces contenus se pose. Ils documentent des pratiques culturelles de notre époque, que ce soit dans les domaines professionnels, scientifiques ou de divertissement. Les informations publiées sur le web sont souvent uniques et ne figurent nulle part ailleurs. D'un point de vue civique, il est nécessaire de documenter les expressions publiques d'entreprises ou de politiciens afin de pouvoir les mettre face à leurs responsabilités en cas de promesses non tenues. Finalement, le monde académique a besoin de pouvoir citer des ressources en ligne sans craindre la disparition de ces liens (*link rot*).

Les raisons pour mettre en place un archivage du web ne manquent donc pas. Il serait tentant de vouloir tout conserver, toute information ayant potentiellement de la valeur pour les historiens du futur. Il est vrai que les technologies de préservation des fichiers numériques ont fait de grands progrès ces dernières années, et que le prix du stockage a chuté. Mais des obstacles persistent : les sites web, avec leur contenu dynamique et leur multitude de formats de fichiers, sont des objets beaucoup plus compliqués à préserver que les « simples » documents bureautiques. Par ailleurs, les volumes sont colossaux.

Un tel contexte nous pousse à nous demander comment évaluer et sélectionner les contenus qui méritent d'être conservés ? Depuis une dizaine d'années, certains pays ont instauré un système de dépôt légal du web ; d'autres initiatives privées ou publiques ont également vu le jour. Mais dans tous les cas, seule une partie du web est sauvegardée. Dans l'absolu, quelle méthode de collecte offre alors le meilleur degré de couverture en matière de présence, de fréquence et de représentativité ? Plus particulièrement, les questions auxquelles nous souhaitons apporter des réponses lors de cette recherche sont les suivantes :

1. Quelles sont les pratiques en matière d'archivage du web (initiatives nationales, commerciales, etc.) ? Lesquelles concernent la Suisse ?

2. Quel est le degré de couverture qui résulte de ces pratiques (présence, fréquence) ?
    - de manière générale (comparaison entre acteur X vs. acteur Y),
    - pour quelques sites exemplaires.

1.2 Attentes

Ce travail de recherche fondamentale, quantitative et exploratoire se divise en plusieurs volets. Nous commençons par examiner les pratiques de l'archivage du web en général, au niveau international, avant d'analyser et comparer les approches de deux acteurs de l'archivage de contenus web suisses :

- La Bibliothèque nationale suisse (BN), qui archive un choix de sites, en fonction de son mandat de collecte des (e-)Helvetica, selon une sélection qualitative ;
- Internet Archive (IA), une initiative américaine à but non lucratif, qui archive toutes sortes de contenus numériques, dont des pages web, en les moissonnant à l'aide d'un crawler.

Ensuite, nous étudions le degré de couverture qui en résulte, sous deux angles différents :

- d'une part en analysant la couverture commune ou différentielle entre des ressources archivées par la BN et par IA respectivement (axe quantitatif),
- d'autre part en étudiant l'exemple de quelques sites suisses (axe qualitatif).

1.3 Définition des concepts

Le **web** est une fonctionnalité d'Internet qui permet la consultation de pages *via* un navigateur, grâce à trois standards clés : URI (uniform resource identifier), HTTP (hypertext transfer protocol) et HTML (hypertext markup language) (Organisation Internationale de Normalisation 2013). Nous distinguons les *pages* web (documents HTML unitaires) des *sites* web (ensembles de pages). D'autres types de fichiers peuvent être intégrés dans les pages web : images, sons, etc. Chaque ressource web possède une URL (Uniform Resource Locator), ce qui permet de la lier à d'autres par des hyperliens. Ces ressources sont stockées dans un ensemble de dossiers et sous-dossiers, organisés par le concepteur du site selon une logique qui lui est propre. Dans l'URL, le chemin d'accès aux sous-dossiers est visible : il s'agit des éléments à la suite du TLD, séparés par des barres obliques.

Un site web est accessible et identifiable par son **nom de domaine** (par exemple hesge.ch ou seco.admin.ch), et c'est ce critère que les organisations d'archivage ont retenu pour identifier les sites.

Le **web suisse** englobe tous les sites qui correspondent à des "Helvetica" selon la définition de la BN. Il s'agit là d'une notion intellectuelle ; techniquement, ces sites sont assez difficilement repérables dans leur globalité (voir section 2.9). C'est pourquoi nous avons appliqué différentes définitions selon les analyses menées (voir section 3.1).

Par **couverture** d'un site web archivé, nous entendons les aspects suivants :

A. présence du site dans la collection,
B. nombre de versions présentes,
C. répartition temporelle de ces différentes versions (fréquence d'archivage),

D. complétude des pages, documents, etc., liés au site, c'est-à-dire si ces éléments sont bel et bien présents dans la version archivée et que les liens pour y accéder fonctionnent.

Les points A à C peuvent être comparés sur des larges ensembles, pour le point D il convient de sélectionner un échantillon.

La **profondeur** de l'archivage d'un site est fonction du nombre de niveaux de sous-dossiers qui sont explorés par le *crawler*. Celui-ci peut par exemple être configuré pour s'arrêter au deuxième sous-niveau, ou au contraire suivre l'ensemble des liens internes jusqu'au dernier sous-niveau.

Archivage du web :

> *"Archiving the web means selecting and capturing internet resources, storing them in Web archives, preserving them and managing sustainable access to the archives. The collecting processes are managed automatically and at regular interval by harvesting software."* (Oury, Poll 2013)

Cette opération d'archivage est souvent appelée *crawl* ou *harvest* (moissonnage). Elle peut être à grande échelle (*bulk harvest,* collecte d'un TLD dans son ensemble) ou sélective (*selective crawl,* collecte d'une liste prédéfinie d'URL).

Tant IA que la BN utilisent le *crawler* Heritrix pour la collecte, et stockent les données selon le format d'archivage du web WARC, tel que normalisé et recommandé par l'IIPC, dont les deux institutions sont membres. Ainsi, les données générées sont de la même nature, facilitant notre préparation et notre analyse des données.

Les **codes de réponse HTTP** (*status code, response code*) indiquent le résultat d'une requête HTTP sous forme d'un nombre à trois chiffres. Quelques exemples fréquents sont 200 (succès de la requête), 301 (redirection permanente), 404 (page non trouvée) ou 503 (erreur du serveur) (Liste des codes HTTP 2019 ; Organisation Internationale de Normalisation 2013). Pour l'internaute, en général, ces codes ne présentent pas beaucoup d'intérêt, à l'exception du 404, assez connu. En revanche, dans le domaine de l'archivage du web, même un code d'erreur constitue une information potentiellement intéressante au sujet d'un site, et il est donc préservé au même titre que le contenu. Dans ce projet, nous nous sommes concentrées sur les enregistrements avec le code HTTP 200.

1.4 Éléments hors périmètre

Dans le cadre de ce travail de recherche, nous n'examinons pas les aspects de conservation des *records* qu'une entreprise se doit de garder dans le cadre de sa gouvernance de données, afin de prouver des droits ou en cas de litige par exemple. Nous ne dressons pas non plus un historique exhaustif de l'archivage du web au niveau international.

Comme nous l'avons déjà dit, nos efforts sont uniquement dirigés sur les réponses positives à une requête HTTP, traduites par un statut de réponse de valeur 200. Les autres codes de statut ne permettant pas d'accéder au contenu des pages web, celles-ci n'engendrent pas la réalisation d'une archive.

Enfin, notons que le web est divisé en différents niveaux selon l'accès possible à son contenu. Nous ne traitons que du web « commun », et plus particulièrement le web de surface. En effet, le web profond (ou *deep web* – à ne pas confondre avec le *dark web*) est constitué

d’informations qui ne sont pas accessibles de premier abord sans le recours à la saisie de mots de passe ou d’outils informatiques particuliers, et ne sont donc pas lisibles par les moteurs de recherche ou les dispositifs de récolte du web pour l’archivage.

2. De l'archivage du web

Dans ce chapitre, nous détaillons l'état de l'art sur notre sujet, en fonction des lectures d'articles, de sites et d'autres publications que nous avons étudiés dans le cadre de notre travail de recherche. En parallèle, nous avons rédigé une bibliographie commentée qui reprend une sélection de ces références (Donius, Hug Buffo 2019).

2.1 Les pages web comme objet d'archivage

Depuis toujours, les supports conservés dans nos institutions patrimoniales évoluent en fonction des avancements technologiques de la société (Aubry 2010). Aux parchemins, manuscrits et livres, se sont ajoutés depuis le XIX^{e} siècle divers supports audiovisuels, puis ceux liés à l'informatique. Mais il s'agissait toujours, dans un premier temps, d'objets physiques (bandes magnétiques, CD-Roms, etc.), à stocker dans des conditions climatiques spécifiques et à consulter sur des appareils dédiés. Plus récemment, les documents nés-numériques font directement l'objet d'un archivage électronique, sans passer par une phase d'existence physique.

Avec l'avènement d'Internet et du World Wide Web (www ; ci-après, web), un type de ressource d'un nouveau genre s'est ajouté au tableau. Les pages web sont en effet d'une structure plus compliquée que d'autres ressources électroniques, tels que les documents bureautiques, constituées d'un item unique (Aubry 2010) ; elles englobent potentiellement des sous-fichiers, du contenu dynamique, des *frames*… Le web dans son ensemble est marqué d'un dynamisme extraordinaire en termes de fluctuation et de renouvellement des contenus, ainsi que d'une pluralité de formats (Crook 2009). La durée de vie d'une page web s'élève souvent à quelques dizaines de jours seulement en moyenne (Kahle 1997 ; Guy 2009), voir à quelques minutes pour les sites d'actualités par exemple.

Internet, dont le web, joue un rôle significatif dans notre société moderne, que ce soit pour la communication, le travail ou le divertissement. On y trouve des informations de toutes sortes, sous différents formats, et souvent uniques. Par conséquent, il est important de documenter les contenus et les usages du web pour les historiens futurs (Costa et al. 2017). Son identité multiple a par ailleurs changé les types et méthodologies de recherche, entre autres en sciences sociales (Gill, Elder 2012). En raison de l'instabilité déjà évoquée des contenus, la citabilité des sources n'est plus garantie, ce qui pose problème pour la recherche académique (Gebeil 2019a). L'UNESCO (2004) a reconnu l'importance de la conservation des contenus numériques pour le patrimoine mondial.

Mais qui est responsable de l'archivage du web ? Au niveau national ou régional, ce sont généralement les bibliothèques patrimoniales qui s'en chargent (voir section 2.4). Pour des domaines bien spécifiques, cela peut aussi être des universités (p.ex. UC Berkeley sur le conflit en Ukraine (Pendse 2016)), des entreprises privées (p.ex. Coca-Cola (Pennock 2013)) ou des chaînes de média (p.ex. la BBC (Smith 2005)). Selon Ullmann et Rösler (2007), les services d'archives devraient aussi se préoccuper des sites web des administrations dont ils dépendent et gérer des questions comme l'évaluation de la valeur archivistique ou les délais de protection. Un colloque du CECO[1] en 2013 est arrivé à la même conclusion :

[1] Centre de coordination pour l'archivage à long terme de documents électroniques, Suisse.

> *« L'archivage de contenus web est encore souvent paré d'une aura extraordinaire, comme si cette tâche ne relevait pas des affaires courantes des archives. Pourtant, celles-ci se doivent d'avoir pour objectif de traiter les contenus web ayant une valeur archivistique dans les mêmes processus ou, du moins partiellement, avec les mêmes outils que toutes les archives numériques. »* (Centre de coordination pour l'archivage à long terme de documents électroniques 2013)

On peut parfois observer, en l'absence d'une politique d'archivage officielle, un « déploiement de formes d'auto-constructions mémorielles » d'associations ou d'autres communautés (Gebeil 2019b). Récemment, on a vu l'entrée sur le marché de l'archivage du web de certains acteurs à but lucratif ; cela peut être considéré comme un signe de maturité du domaine (Pennock 2013).

2.2 Technologie

La technologie qui permet de moissonner les sites à archiver fonctionne comme l'indexation pratiquée par les moteurs de recherche : un *crawler* est alimenté par une *seed list*, c'est-à-dire des URL de départ. Il envoie une requête HTTP pour récolter le contenu de la première URL. S'il y a des liens, il les ajoute à la liste des pages à visiter. Selon la configuration du *crawler*, le moissonnage est plus ou moins profond : il s'arrêtera par exemple au troisième niveau de sous-page d'un site.

De grands pans du web sont inaccessibles par cette méthode :

- Certains sites ne sont pas liés à d'autres par des liens, ou encore, un mot de passe est requis pour accéder à certains contenus ; c'est ce qu'on nomme le *"deep web*".
- Les outils de moissonnage rencontrent des limitations face à des pages générées dynamiquement ou à certains formats de contenu (p. ex. Flash).
- Le webmaster d'un site a la possibilité d'indiquer au *crawler* les ressources qu'il ne souhaite pas voir moissonnées, en plaçant un fichier nommé robots.txt à la racine du site. Or, le *crawler* peut en tenir compte ou pas.
- etc.

Pour maintenir les archives lisibles à travers le temps, l'enregistrement et la sauvegarde des sites répondent à des standards techniques permettant un échange et la visualisation des contenus. Ces outils ont souvent été initiés par l'acteur précurseur de l'archivage du web, Internet Archive (IA ; voir section 2.3). Dans le cadre de l'IIPC (voir section 2.6), IA a collaboré avec d'autres acteurs à l'élaboration d'un format d'archivage, le WARC, qui fait désormais l'objet d'une norme internationale : l'ISO 28500 (Oury, Poll 2013 ; Organisation Internationale de Normalisation 2017). Le CDX, quant à lui, est un format de fichier dans lequel chaque ligne résume un document du web. Il permet par ailleurs la création d'index des éléments archivés et facilite la recherche de ceux-ci (IIPC 2015).

En raison de la fluidité des contenus, les copies archivées des pages ne sont pas forcément le reflet exact du *live web*. Il arrive par exemple que les images présentes sur une page HTML soient moissonnées à un autre moment que le contenu textuel, ce qui peut générer une incohérence entre l'image et sa légende. Brügger (2009) énonce que l'archive du web est une « reconstruction subjective, activement créée » ; selon lui, plutôt que d'ancrer une version archivée à un moment précis, il faudrait parler d'un continuum temporel. D'après Musiani et collègues (2019), l'archive du web est « un objet singulier, interactif, fluide et non figé ». Gebeil

(2019a) la qualifie de *re-born digital heritage* : la page web est déjà le résultat d'une médiation de différents éléments, que le processus d'archivage répète.

2.3 Brewster Kahle et Internet Archive

Un pionnier dans les démarches d'archivage du web est l'américain Brewster Kahle, qui rêve de créer « une deuxième bibliothèque d'Alexandrie » (Lepore 2015 ; Burns 2019) et de donner « un accès universel à tout le savoir du monde » (Internet Archive 2019a). Il a fondé en 1996 Internet Archive dans le but d'offrir à une large communauté de chercheurs et historiens un accès permanent à des collections sous forme numérique. La majeure partie des contenus est constituée de pages web, mais des documents audio ou vidéo, ainsi que des logiciels sont également archivés (Internet Archive 2019a). Basée à San Francisco, IA a ses locaux dans une ancienne église qui ressemble à un temple grec – et donc au logo de la société à but non lucratif (Lepore 2015).

Brewster Kahle a obtenu son diplôme d'informaticien au MIT en 1982 et est le créateur, entre autres, des entreprises WAIS (analyse des contenus de serveurs Internet et création de listes dans le but d'améliorer la recherche, à la période pré-web ; vendu à AmericaOnline en 1996) et Alexa Internet (logiciel permettant de tracer le trafic sur Internet et de recommander des sites ; vendu à Amazon en 1999) (Hardy 2009). La vente de ces deux entreprises lui a fourni le capital nécessaire pour lancer IA (Chen 2006 ; Hardy 2009). L'organisation, qui emploie environ 150 collaborateurs, est aujourd'hui principalement financée par des dons, auxquels appelle une rubrique dédiée sur le site : *"If everyone who uses the Internet Archive donates just $ 5, we can keep offering these services for free and ad-free"* (Internet Archive 2019b). Des fondations américaines donnent également des sommes importantes, comme par exemple le Pineapple Fund, qui, fin 2017, a offert 1 million de dollars sous forme de Bitcoins (Barrett 2017). En complément, IA monnaie les services fournis aux différentes bibliothèques nationales avec lesquelles il collabore pour des moissonnages de leur domaine de premier niveau *via* son service Archive-It (Archive-It 2019a ; 2019b), ou pour la numérisation de leurs collections de livres (Hardy 2009).

Selon Brewster Kahle, les documents archivés doivent être utilisés et en accès ouvert afin de garantir leur conservation – les enfermer dans un dépôt, aussi sécurisé soit-il, ne serait pas la bonne voie (Kahle cité dans Minard (2013)). C'est pourquoi la majeure partie des collections d'IA sont accessibles en ligne. Pour les archives du web, l'accès se fait par la Wayback Machine, interface développée dès 2001 (Edwards 2004) : après saisie de l'URL recherchée, une *timeline* permet de visualiser la chronologie des différentes versions archivées du site correspondant et de passer de l'une à l'autre. En revanche, il n'existe pas de recherche plein texte dans les contenus d'IA.

Un fait intéressant : Brewster Kahle, qui a pourtant nommé son « bébé » Internet *Archive*, se décrit lui-même comme "Digital *Librarian*" (Minard 2013), et c'est également le terme *library* qui est utilisé dans le site. Il s'agit effectivement d'une bibliothèque, car IA assemble des collections d'origines diverses, contrairement à un service d'archives qui se préoccupe des documents générés par un producteur spécifique (Monks-Leeson 2011). Le point de vue contraire est défendu par Edwards (2004) : pour lui, IA n'effectue pas certaines tâches classiques d'une bibliothèque, telles que la description bibliographique des documents ou la sélection en fonction d'une politique des collections.

2.4 Pratique de la Bibliothèque nationale suisse

De nombreux pays disposent d'une politique d'archivage de sites web considérés comme patrimoniaux par des institutions nationales. Les approches sont diverses : tantôt on cherche l'exhaustivité d'un ccTLD, tantôt on procède par sélection, tout en essayant de garantir la représentativité (Brown 2006). Une liste très complète de ces différentes initiatives, instaurée après une enquête internationale auprès des acteurs de l'archivage du web (Gomes et al. 2011), et régulièrement mise à jour, est accessible sur Wikipédia (List of Web archiving initiatives 2020). Pour les institutions patrimoniales, il est nécessaire d'être proactives en matière d'archivage de ce type de contenu ; ce ne sont pas les producteurs (dont notamment l'administration publique) qui s'en préoccuperont (Shein 2016 ; Crook 2009).

En Suisse, c'est la Bibliothèque nationale (BN) qui s'en charge, en vertu de son mandat de conservation des "Helvetica", c'est-à-dire de toutes les publications suisses / sur la Suisse / par des Suisses, tous supports confondus (Bibliothèque nationale suisse 2018a). Elle a lancé la réflexion en matière d'archivage du web suisse en 2001 (Signori 2019a), en collaboration avec des institutions cantonales. Encore aujourd'hui, ces partenaires proposent les sites web à conserver, selon des critères communs (Signori 2019b). La sélection de sites prend en effet plus de temps que ce qu'on pense (Crook 2009), et il faut bien connaître le contexte d'une région pour pouvoir choisir des sites réellement représentatifs. La BN s'occupe des aspects techniques et de la coordination. L'accès aux collections n'est pas possible en ligne, il faut se rendre dans les locaux d'une des bibliothèques partenaires.

Le fonctionnement technique du système de la BN est décrit dans la « Notice Archivage » (Locher 2015) : le *crawler* Heritrix est configuré pour collecter la totalité des sous-répertoires et documents dépendants de l'URL de départ, à moins qu'il n'y ait une restriction technique du côté du site web. Le processus de moissonnage est piloté et surveillé via une interface web.

Différents contrôles qualité sont effectués manuellement : nombre de documents présents, structure des répertoires, intégrité d'une chaîne de liens… Il y a également, au fur et à mesure, une comparaison visuelle automatisée entre le site web *live* et ce qui vient d'être moissonné. En cas de dissemblance, le processus est répété après adaptation de la configuration. Le fonctionnement des éléments dynamiques (recherche, calendrier, diaporama, script…) est également contrôlé, mais ne conduit pas obligatoirement au rejet du site en cas de défaut. Les bibliothécaires responsables des collections prennent la décision finale (Locher 2015).

Par la suite, un identifiant pérenne (URN) est attribué à l'archive et les métadonnées sont intégrées aux catalogues. Le système d'archivage est conforme au modèle OAIS (Open Archival Information System), modèle logique de référence dans le monde des bibliothèques et archives numériques (Locher 2015).

L'approche sélective de la BN (par opposition à une collecte large comme elle est pratiquée par exemple en France) est surtout due à l'absence d'un dépôt légal national dans notre pays (Beausire 2015). Un moissonnage de la totalité du domaine .ch n'est toutefois pas exclu à tout jamais ; pour l'instant, les ressources ne sont pas disponibles (Locher 2015).

2.5 Pratique d'Internet Archive

IA n'effectue *a priori* pas une sélection, mais moissonne tout ce qu'il peut. Contrairement à un moteur de recherche généraliste, qui procède à l'indexation du web en faisant tourner un grand

nombre de *crawlers* simultanément selon les mêmes configurations, il y a chez IA une grande variété de collections alimentées par différents *crawlers* en parallèle, selon les besoins divers (Leetaru 2016).

Initialement, IA tenait compte des fichiers robots.txt et renonçait à la collecte des pages indiquées par ce protocole d'exclusion des robots. Les anciennes versions étaient même effacées des archives, si un fichier robots.txt était découvert lors d'un nouveau passage par un site. Cette politique respectueuse des souhaits des webmasters a été abandonnée dès avril 2017 (Graham 2017). Aujourd'hui, toutes les ressources sont moissonnées et archivées. Il est néanmoins possible de demander à IA par e-mail de supprimer un contenu et/ou de le rendre introuvable dans la Wayback Machine.

2.6 Collaborations multiples

Depuis les débuts de l'archivage du web, les différentes institutions patrimoniales nationales ont opté pour la collaboration afin de partager les efforts. Ce fut le cas du projet NEDLIB (Networked European Deposit Library), financé par l'Union Européenne, qui regroupe des partenaires venant d'Allemagne, de Finlande, de France, d'Italie, de Norvège, des Pays-Bas, du Portugal et de Suisse, ou du projet NWA (Nordic Web Archive) des bibliothèques nationales de l'Europe du nord. Les expériences de ces démarches étant précieuses, elles furent reprises ensuite au sein de l'IIPC (Hakala 2004).

Le Consortium international pour la préservation de l'Internet (International Internet Preservation Consortium, IIPC) a été créé en 2003 par une dizaine de bibliothèques nationales européennes et anglo-saxonnes. Au début, il s'agissait d'un groupement d'experts « technocrates » qui se réunissaient (souvent virtuellement) pour débattre de questions techniques (Illien 2011). Avec la maturité grandissante des différentes initiatives et l'association de toujours plus d'acteurs (actuellement, les membres proviennent de près de 50 pays), le consortium couvre aujourd'hui des sujets divers comme la standardisation des formats, le lobbying auprès des décideurs et les solutions d'accès. IA et la BN sont tous deux membres de l'IIPC.

D'autres collaborations internationales existent pour des sujets spécifiques. À titre d'exemple, nous pouvons citer RESAW (REsearch infrastructure for the Study of Archived Web materials), un réseau d'institutions européennes dédié depuis 2012 à l'exploitation académique transversale des données archivées et financé partiellement par le programme Horizon 2020 de l'Union Européenne (RESAW 2019).

2.7 Aspects légaux

Le cadre légal varie d'un pays à l'autre : parfois, il existe un dépôt légal qui couvre également les sites web, ce qui facilite évidemment les démarches pour leur archivage. Dans une telle loi (mais également dans les documents internes aux institutions qui archivent), il convient d'utiliser une définition assez large de l'objet de collection, afin de couvrir les évolutions techniques à venir (Illien 2008).

Contrairement aux publications imprimées, il n'est pas possible de contacter les maisons d'édition pour leur demander de faire parvenir à la bibliothèque, dans le cadre du dépôt légal, une copie de tout ouvrage publié ; et cette disposition légale ne devrait pas contenir une

quelconque obligation pour l'éditeur, mais le droit pour l'institution patrimoniale de moissonner ce qui relève de sa compétence (Hakala 2004).

L'archivage de contenus du web entre en général en conflit avec les législations sur le droit d'auteur et sur la protection des données (Brunner 2014 ; Berčič 2005). Le fait qu'il s'agisse d'activités potentiellement transnationales ne facilite pas la question (Illien 2011). C'est pourquoi les institutions patrimoniales limitent le plus souvent l'accès à leurs collections web : la consultation n'est possible que dans des salles dédiées (Bibliothèque nationale suisse 2019 ; Aubry 2010). Deux exceptions qui confirment la règle sont IA et les archives portugaises du web, qui proposent un large accès en ligne (Schafer et al. 2016).

2.8 Évaluation de la couverture de l'archivage web

Afin de pouvoir justifier leurs activités face aux bailleurs de fonds, les bibliothèques doivent être en mesure de les chiffrer et de calculer leur performance ; c'est le cas également pour les programmes d'archivage du web (Oury, Poll 2013). Les institutions membres de l'IIPC ont bien compris la nécessité de disposer d'indicateurs communs pour l'évaluation quantitative et qualitative de leurs activités. C'est pourquoi ils ont élaboré ensemble, depuis 2009, un rapport technique sur les *Statistiques et indicateurs de qualité pour l'archivage du web*, adopté par l'ISO en décembre 2013. Ils ont choisi de ne pas créer une norme, document d'une forme plus contraignante, afin de pouvoir le publier plus vite (Oury, Poll 2013).

Les données collectées à des fins d'analyse doivent être :

- appropriées pour l'évaluation du service,
- comparables entre les services et à travers le temps,
- un support pour la gestion interne du service,
- favoriser la reconnaissance de la valeur du service (Oury, Poll 2013).

Au niveau quantitatif, le rapport technique propose des indicateurs sur la taille des archives (nombre d'URL cibles, volume en octets…), sur le contenu (distribution géographique, formats de fichiers…), sur l'utilisation des archives (accès en salle de lecture…), sur les mesures de conservation à long terme (taux de réplication sur un site distant, stratégies de migration…) et sur les coûts.

Au niveau qualitatif, il s'agit d'évaluer l'efficience et l'efficacité des démarches entreprises par rapport aux objectifs de l'institution. On trouve des indicateurs tels que le pourcentage de ressources cataloguées et/ou indexées, la couverture du périmètre inscrit dans la mission ou le taux de ressources archivées ayant disparu du *live web* (Oury, Poll 2013).

2.9 Définition d'un « web national »

La conception du web ne connaît *a priori* pas de frontières étatiques. Le fait de vouloir créer au moment de son archivage des silos nationaux (Schafer et al. 2016) semble quelque peu contraire à cet esprit transnational. Cela rajoute d'ailleurs également des obstacles pour des projets de recherche concernant toute une région englobant plusieurs pays (Gebeil 2019b). Mais cette approche est notamment due à l'organisation des différentes institutions collectrices, qui doivent se conformer à leur cadre législatif et à leur mission (Illien 2011).

Hakala (2004) indique qu'il n'est pas suffisant de se limiter à un ccTLD spécifique ; des sites en .com ou .org font également partie du « web national » d'un pays donné. Il liste quelques

méthodes pour obtenir d'autres URL : par l'organisation qui attribue les noms de domaine (cela se révèle souvent difficile, l'information étant considérée comme propriétaire par cette organisation) ; par des fournisseurs d'accès (en Finlande, les chercheurs n'ont réussi à obtenir les informations que de 2 fournisseurs sur 10) ; par une autre organisation qui dispose déjà d'une telle liste, assemblée par exemple dans le cadre d'une recherche (Hakala et ses collègues ont ainsi pu obtenir 60'000 noms de domaines, dont beaucoup avec un TLD autre que .fi) ; selon le contexte, des méthodes linguistiques permettent de sélectionner les pages écrites dans une certaine langue (cela fonctionne bien pour le web finlandais, mais pas pour des langues telles que l'anglais, l'espagnol ou l'allemand, qui sont pratiquées dans de nombreux pays).

En 2004, Hakala estimait qu'environ 40 % du web suédois et finlandais se situait en dehors des ccTLD .se et .fi respectivement, et il supposait que ce nombre allait augmenter avec l'introduction de nouveaux TLD génériques. Outre les domaines de type internationaux tels que .com, .net ou .org, les propriétaires de sites aiment bien aussi choisir des extensions qui jouent sur les significations implicites, p.ex. .nu pour les sites suédois (Arvidson 2002) – *"nu"* signifiant « maintenant » en suédois, mais aussi en danois et en néerlandais – ou .li pour les sites suisses germanophones – *"-li"* étant la terminaison indiquant le diminutif d'un mot en suisse allemand. Cela pourrait d'ailleurs créer des difficultés si Niue (petit pays insulaire en Océanie) ou le Liechtenstein décidaient un jour de vouloir collecter leur propre web national… Cette problématique de faux positifs pour un ccTLD n'a, à notre connaissance, pas encore été étudiée dans la littérature scientifique.

Les professionnels des archives du web ont donc besoin d'une définition de leur « espace virtuel national » afin d'étendre le champ de leur collecte au-delà d'un ccTLD spécifique. Pour la France, c'est la Loi relative au droit d'auteur et aux droits voisins dans la société de l'information (loi DADVSI) du 1er août 2006 et son décret d'application du 19 décembre 2011 qui la donne : « il s'agit tout d'abord des sites hébergés sur des domaines de haut niveau français (.fr, .paris, .re pour l'île de la Réunion, etc.) ; et/ou des sites dans un nom de domaine enregistré par une personne domiciliée en France ; et/ou enfin des sites produits sur le territoire français » (Bonnel, Oury 2014).

Pour caractériser le web portugais, ont été définies comme « pages ayant un intérêt culturel et sociologique pour le peuple du Portugal », en plus de celles hébergées sur un domaine du ccTLD .pt, les pages web de langue portugaise, hébergées sur un domaine des gTLD .com, .net, .org ou .tv, avec au moins un lien y pointant depuis un domaine du ccTLD .pt (Gomes, Silva 2005). Les liens entrants jouent donc un rôle important dans l'évaluation quant à l'appartenance à un « espace national ».

Vlcek (2008) a poussé plus loin encore ce raisonnement en ce qui concerne le web tchèque. Il a développé un module qui s'intègre au *crawler* Heritrix. Programmé en Java, WebAnalyzer examine les pages en dehors du ccTLD .cz mais atteint par un lien depuis ce domaine. Il attribue des points selon des critères paramétrables, par exemple si le texte contient des adresses e-mail finissant par .cz, si des mots d'une liste définie y figurent, si l'attribut *lang* dans le code source se réfère à la langue tchèque ou encore si l'adresse IP est localisée en République tchèque. Lorsqu'une page web totalise un certain nombre de points, elle est considérée comme tchèque et par conséquent archivée.

Quant à la Suisse, la BN a pour mission de conserver les “Helvetica”, donc tous les documents ayant un rapport à la Suisse, selon la *Loi sur la Bibliothèque nationale (LBNS)* du 18 décembre 1992, art. 3.1 :

> *« La Bibliothèque nationale collectionne les informations imprimées ou conservées sur d’autres supports que le papier, qui :*
>
> - *paraissent en Suisse ;*
> - *se rapportent à la Suisse, à ses ressortissants ou à ses habitants ou*
> - *sont créés, en partie ou en totalité, par des auteurs suisses ou par des auteurs étrangers liés à la Suisse. »*
>
> (Confédération suisse 1992)

Potentiellement, des sites de n’importe quel TLD ou de n’importe quelle localisation géographique peuvent tomber sous cette définition. La BN n’a pas recours à des moyens techniques pour identifier les sites suisses, étant donné qu’elle a opté pour une approche de sélection manuelle des URL à archiver : « Elle met l'accent sur les sites web patrimoniaux qui ont un fort lien avec la Suisse et qui sont accessibles librement : sites web sur les cantons et les communes, domaines spécifiques tels que sciences sociales ou littérature suisse. » (Bibliothèque nationale suisse 2018a).

Le registre du TLD suisse par excellence .ch est tenu par Switch sur un mandat de l’Office fédéral de la communication (OFCOM 2016). C’est aussi cet office qui gère directement le gTLD .swiss, disponible depuis 2015 (OFCOM 2018). Les données concernant ces deux TLD sont donc considérées comme des archives publiques (Archives fédérales 2017). Fin 2018, il y avait un peu plus de 20’000 noms de domaines enregistrés sous le gTLD .swiss (OFCOM 2019) et presque 2,2 Mio sous .ch (Switch 2019a). Ce dernier représente plus de 55 % de tous les noms de domaine enregistrés par des personnes ayant une adresse en Suisse. Néanmoins, nous savons déjà que cela ne correspond pas à la totalité du “web suisse” – le TLD le deuxième plus populaire est le .com avec près de 25 %, mais nous trouvons aussi du .org, .de, .info, .uk, etc. (Switch 2019b).

3. Méthodologie

L'objectif de ce projet de recherche est l'évaluation de la couverture de l'archivage du web suisse. Nous avons établi qu'il nous fallait, pour y parvenir, des informations au sujet des moissonnages des sites et pages web dans le but de leur archivage. Tout en menant nos études exploratoires quant à la méthode de collecte de données, mais également du type de (méta-)données qu'il nous serait possible d'obtenir, nous avons identifié ce que nous entendions par "web suisse".

3.1 Définition du « web suisse »

Selon la définition de la LBNS et de l'ordonnance qui la spécifie (OBNS), les "Helvetica" sont tous les documents qui paraissent en Suisse ; qui se rapportent à la Suisse, à ses ressortissants ou à ses habitants ; ou qui sont créés, en partie ou en totalité, par des auteurs suisses ou liés à la Suisse. Cette définition est assez large, et déjà pour les supports physiques il n'est pas évident de tenir une liste exhaustive de toutes les œuvres qui y correspondent. Pour ces documents immatériels et volatiles que sont les sites web, cela est encore plus difficile. La section 2.9 du présent rapport en donne quelques détails.

Pour des raisons de faisabilité, nous avons adopté plusieurs approches dans la définition du web suisse, selon nos axes de recherche.

3.1.1 Pour la comparaison IA vs. BN : les sites avec le ccTLD .ch

En effet, pour notre première analyse quantitative, aucun des outils disponibles auprès d'IA pour accéder aux sites archivés et leurs métadonnées ne permet de qualifier la « nationalité » d'un site web, et ceci pour les mêmes raisons que celles évoquées dans la section 2.9. Il aurait certes été possible d'effectuer une analyse poussée sur les données pour déterminer la « suissitude » des sites, basée par exemple sur des mots-clefs ou sur l'existence de liens entre différents sites. Mais nous ne disposions ni des capacités techniques, ni des compétences, ni du temps requis pour capter, analyser et explorer le volume de données correspondant et les différentes méthodes de sondage. Nous avons par conséquent adopté une approche plus pragmatique, en tenant compte du fait que Switch est le *registrar* officiel pour les deux TLD .ch et .swiss. Ces deux TLD sont considérés officiellement comme ceux de la Suisse, et donc du web national. Cependant, le nombre de noms de domaines enregistrés est très inégal : alors que .swiss ne comptabilise que 20'000 noms de domaines en 2019, le nombre de .ch est de presque 2,2 Mio (> 99 % du total). Nous avons par conséquent concentré nos efforts sur cette dernière extension dans notre analyse chez IA.

3.1.2 Pour la comparaison BN vs. IA : les sites archivés par la BN

Pour notre seconde analyse quantitative, basée en premier lieu sur le corpus de sites archivés par la BN, il a été plus évident de déterminer la définition du web suisse que nous allions utiliser : il s'agit de la même que celle appliquée par l'institution. Mais surtout, le nombre de sites web archivés, beaucoup plus restreint que chez IA, nous permet de considérer la totalité de ses références.

3.1.3 Pour la comparaison qualitative : un échantillon de sites

Enfin, pour notre analyse qualitative, nous nous sommes basées sur la définition de "Helvetica" de la BN pour dresser une liste de sites que nous connaissons par le biais de notre pratique professionnelle ou personnelle.

3.2 Définition des items

Notre exploration des méthodes de collecte nous a permis d'identifier trois types de ressources qui allaient pouvoir servir notre recherche.

Ainsi, nous avons constaté que tant pour la BN que pour IA, un item consiste en une page web, dépendant elle-même d'un site web. Alors qu'il peut sembler intéressant de comparer la représentativité de chaque institution en mesurant le nombre de **pages** archivées pour chaque site par chacune des institutions, nous avons aussi conscience de la volatilité de ce nombre. En effet, chaque propriétaire de site apporte des modifications à la structure de son site web en fonction de ses besoins propres, de la nature du site (un blog régulièrement alimenté verra son nombre de pages augmenter plus rapidement qu'un site institutionnel), du temps qu'il a à y consacrer, entre autres raisons. À partir de ce constat, nous avons choisi de nous concentrer principalement sur les **sites** archivés comme objet d'analyse, tout en considérant les points suivants :

1. L'existence d'un site web est instable aussi, et un site peut tout à fait disparaître du jour au lendemain. Néanmoins, dans le cadre de ce projet de recherche, il ne s'agit pas d'évaluer son existence actuelle, mais s'il a fait l'objet d'une opération de conservation à un moment de sa vie.
2. Un site est accessible *via* un nom de domaine. Un nom de domaine peut être composé de plusieurs niveaux ; cependant, seul le premier niveau est enregistré auprès des instances régulatrices et comptabilisé dans les statistiques de celles-ci. Il est possible de créer autant de niveaux de noms de domaines que souhaité, menant chacun à un site web distinct. Pour les mêmes raisons de volatilité que pour les pages, nous avons exclu les sites archivés qui ne correspondent pas à un premier niveau de nom de domaine.
3. Tout nom de domaine peut servir pour une redirection vers un autre site. Nous avons écarté ces cas en excluant les statuts de réponse correspondant à une redirection (codes 3xx).

Les considérations ci-dessus s'appliquent principalement à nos analyses quantitatives. Pour la partie qualitative, nous avons souhaité explorer l'archivage d'un échantillon assez varié et avons donc retenu des items qui ne répondent pas forcément à ces critères « techniques ».

3.3 Définition des ensembles de données

Afin de mener à bien notre recherche, nous avons défini différents ensembles de données en fonction des analyses que nous souhaitions effectuer et de la méthode de sélection. Les données utilisées pour notre analyse quantitative (voir Tableau 1) sont le fruit d'une automatisation de leur collecte (par nous-même ou par l'institution concernée). Pour notre analyse qualitative, en revanche, nous avons effectué une collecte manuelle des données (voir chapitre 5).

Tableau 1 : Ensembles de données pour l'analyse quantitative

Nom	Format	Description
BN1	CDX	Données CDX de la première capture de l'ensemble des sites "Helvetica" archivés par la BN, selon liste extraite des fichiers XML
BN2	CDX	Données CDX de la première capture des sites .ch archivés par la BN, selon liste extraite des fichiers XML (voir 3.8.2)
IA1	CDX	Données CDX de la première capture des sites .ch archivés par IA
IA2	CDX	Données CDX de la première capture par IA des sites archivés par la BN

3.4 Récolte des données de la BN pour l'analyse quantitative

À la BN, au moins une dizaine de métadonnées (Bibliothèque nationale suisse 2018c) pour les objets à archiver sont saisies[2] dès le début du processus de collecte, dans le formulaire d'annonce des sites identifiés pour l'archivage. Ces informations sont ensuite reprises pour figurer dans le catalogue Helveticat, mais également dans une mesure plus réduite dans le catalogue e-Helvetica Access (le titre, l'URL et la classification Dewey). Bien qu'il existe un lien entre les deux catalogues, ils ne partagent pas la même base de données. Nous les avons donc analysés indépendamment afin de tenter de récupérer les données nécessaires à nos analyses.

3.4.1 Helveticat (catalogue global)

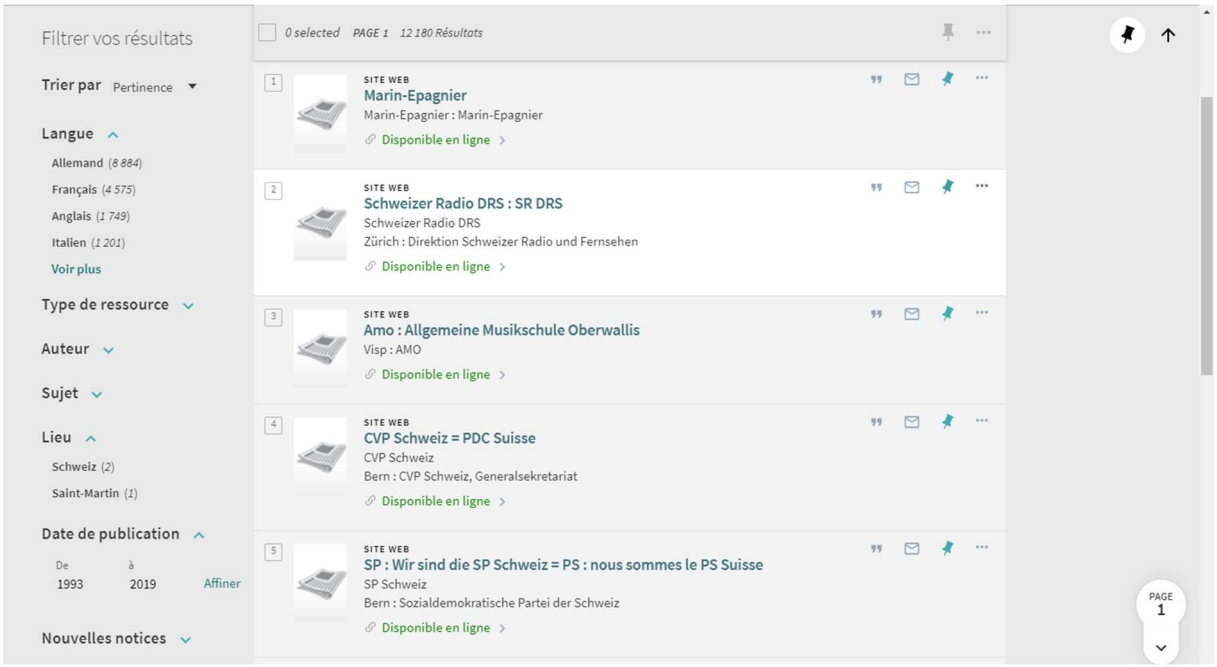
Le catalogue en ligne Helveticat[3] contient les informations sur les "Helvetica" préservés à la BN, qu'ils soient analogiques ou numériques, y compris la collection des sites web patrimoniaux.

Au moment de notre analyse, le catalogue est accessible *via* l'outil de découverte Primo, produit de la compagnie Ex Libris. L'outil propose un moteur de recherche par mot-clef et il est possible de sélectionner le type de document recherché. D'après le *Mode d'emploi pour la recherche de sites web* (Bibliothèque nationale suisse 2018b), le mot-clef (code d'extraction) "xwas" permet d'accéder aux Archives Web Suisse. Combiné au type de document "Sites Web", on trouve 12'180 références en date du 8 janvier 2020.

Les résultats sont affichés sur plusieurs pages et des filtres peuvent être appliqués. Comme l'outil de découverte est avant tout dédié aux références bibliographiques, l'aperçu des résultats fournit les informations attribuées suivantes : le titre, la publication, parfois l'auteur, ainsi que l'information « disponible en ligne », dans la mesure où il s'agit de ressources numériques.

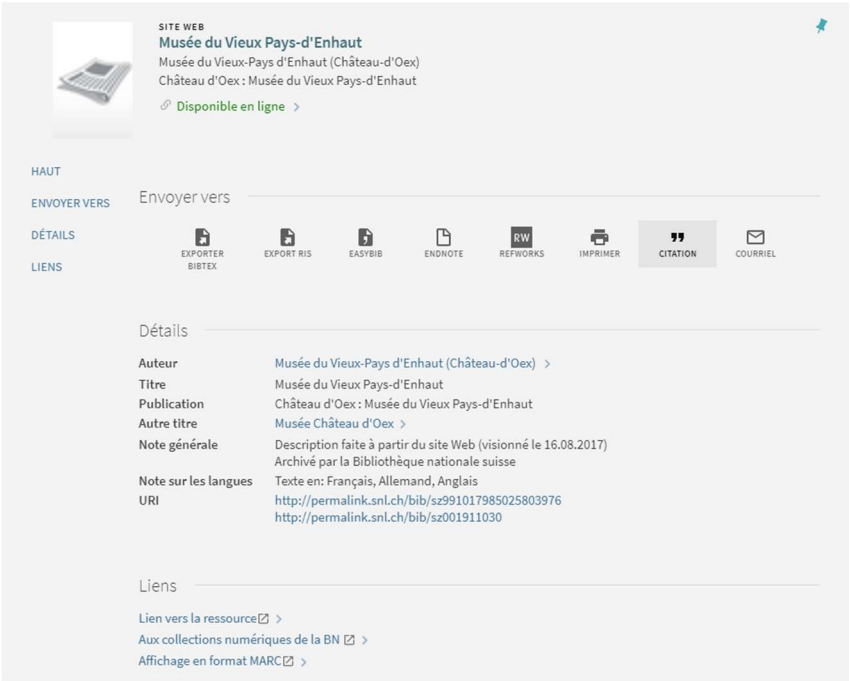
[2] Les métadonnées dont la saisie est obligatoire sont : l'URL, le titre, le nom du producteur, le lieu du producteur, le canton du producteur, le pays du producteur, la langue du site, la classification Dewey, la fréquence de collecte.
[3] https://www.helveticat.ch/discovery/search?vid=41SNL_51_INST:helveticat

Figure 1 : Capture d'écran d'une liste de notices résultant d'une recherche



Le détail des notices est plus complet, avec davantage de métadonnées, et propose leur export vers différents formats de logiciels de gestion de références bibliographiques. L'impression, la citation ou encore l'envoi par e-mail sont également possibles. Depuis la section Liens, et selon la disponibilité, on peut accéder à la version actuelle en ligne de la ressource, à d'anciennes URL qui ne sont plus actives, au catalogue e-Helvetica Access (voir 3.4.2), et enfin à la notice bibliographique de la ressource au format MARC21. Ce format de catalogage standardisé permettrait l'analyse automatisée du contenu des champs.

Figure 2 : Capture d'écran du détail d'une notice de site web dans le catalogue Helveticat



Nous avons envisagé dans un premier temps de récupérer ces informations de catalogage pour mener notre analyse, ou tout du moins pour obtenir une liste exhaustive des sites archivés. Or, les dates de moissonnage des sites n'y figurent pas, alors qu'il s'agissait d'une information requise pour nos analyses. Les pages contiennent par ailleurs des éléments de navigation dynamiques, qui rendent impossible la récupération automatisée du contenu. L'URL de requête présente bien la possibilité d'intégrer des filtres qui outrepassent certains de ces éléments bloquants, mais pas tous.

Figure 3 : Capture d'écran du détail
d'une notice bibliographique en MARC21

```
leader  02159nai a22005297a 4500
001     991017981423903976
003     Sz
005     20191114172036.0
007     cr
008     171003cuuuu9999sz uu wss     0    2ger d
024     7#$ahttp://permalink.snl.ch/bib/sz991017981423903976 $2uri
024     7#$ahttp://permalink.snl.ch/bib/sz001602557 $2uri
035     ##$aoai:nb.ingest:alma-bel-156685
035     ##$a(Sz)991017981423903976
035     ##$a(Sz)001602557
040     ##$aSz $cSz
041     1#$ager $afre
082     74$a320 $222sdnb
245     00$aSP : $bWir sind die SP Schweiz = PS : nous sommes le PS Suisse
246     1#$iFranz. Titel <2009>: $aParti socialiste suisse
246     1#$iFranz. Titel <2009-2014> $aPS Suisse
247     10$aSozialdemokratische Partei der Schweiz $f<2009>
247     10$aSP Schweiz $f<2009-2014>
260     ##$aBern : $bSozialdemokratische Partei der Schweiz
336     ##$btxt $2rdacontent
337     ##$bc $2rdamedia
338     ##$bcr $2rdacarrier
500     ##$aArchiviert durch die Schweizerische Nationalbibliothek
500     ##$aBeschreibung basiert auf Website (gesehen am 24.04.2018)
516     ##$aHTML
546     ##$aText in: Deutsch, Französisch
659     ##$aPOLITISCHE PARTEIEN (IDEOLOGIE UND ZIELE) $xsozarch
659     ##$aSOZIALISTISCHE PARTEIEN + SOZIALISMUS (POLITIK) $xsozarch
659     ##$aNationalratswahlen $xwahlen2007 $xwahlen2011 $xwahlen2015
659     ##$aPartei $xchpart
710     2#$aSP Schweiz $0(DE-588)13728-5 $2gnd
856     40$uhttps://www.sp-ps.ch
```

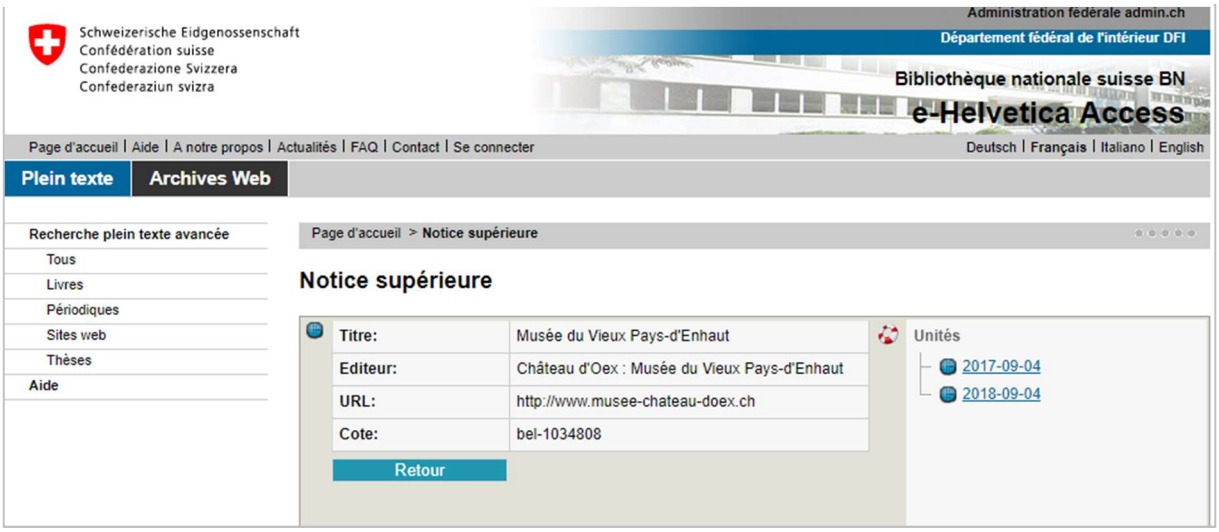
Une autre piste aurait été de recourir à la fonction de sélection multiple disponible au-dessus des résultats, et permettant des exports. Cependant le nombre de résultats est limité à 50 et ne permet pas non plus de récupérer l'entier des résultats.

3.4.2 e-Helvetica Access

e-Helvetica Access[4] est l'interface d'accès aux ressources numériques de la BN. Nos premières prospections sur ce catalogue au début du mois de mai 2019 nous paraissaient prometteuses : la navigation sur le site était moins sujette à des éléments de navigation dynamique, et aurait certainement permis de rassembler une grande partie des données dont nous avions besoin pour nos analyses (voir Figure 4).

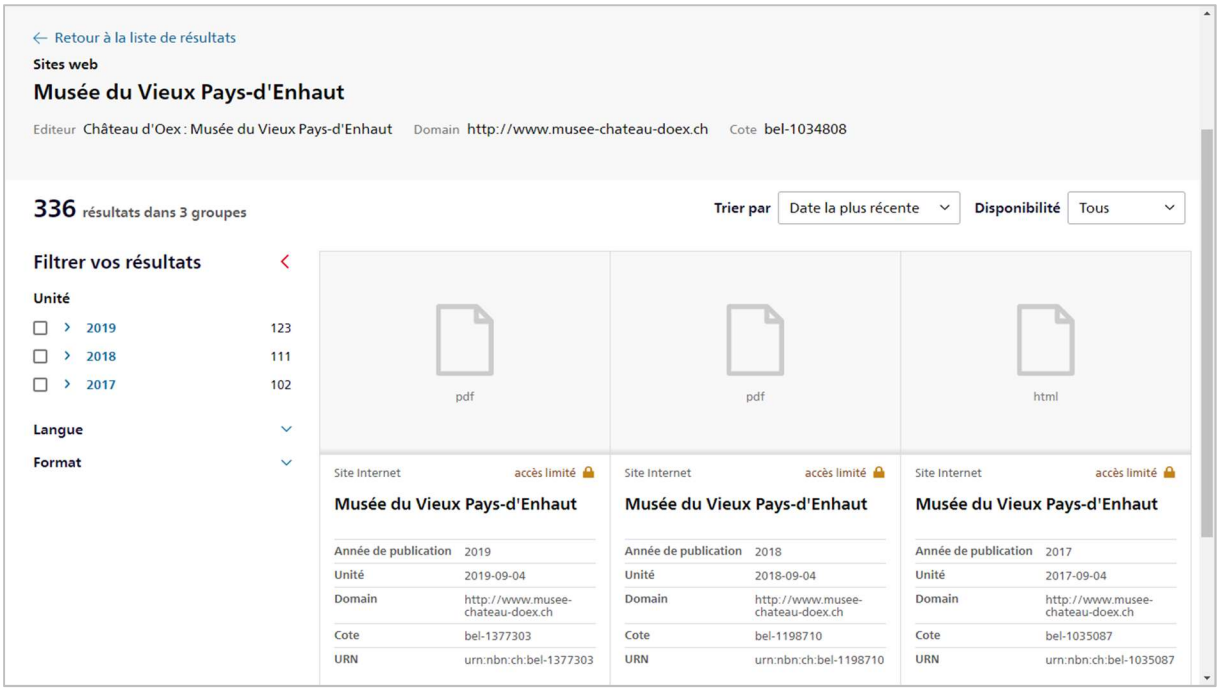
[4] https://www.e-helvetica.nb.admin.ch/

Figure 4 : Capture d'écran de la notice supérieure d'un site web archivé sur e-Helvetica Access en date du 9 mai 2019



Seulement, lorsque nous nous sommes de nouveau rendues sur le site au début du mois de juillet 2019 pour configurer notre moissonnage, nous avons eu la désagréable surprise de constater que la technologie utilisée avait changé. D'apparence plus moderne, la navigation automatisée s'en retrouve complexifiée (voir Figure 5).

Figure 5 : Capture d'écran du résultat pour une notice générale d'un site web archivé sur e-Helvetica Access après le 7 juillet 2019



Face à ce nouveau paradigme, nous avons quand même analysé la nouvelle présentation des sites archivés et tenté d'en dégager une méthode de collecte. Nous avons notamment constaté que l'URL des pages de résultat est paramétrable avec des filtres – de manière

similaire au catalogue Helveticat – avec l'année de publication, une date d'un moissonnage spécifique ou la cote d'un site. Par exemple :

> https://www.e-helvetica.nb.admin.ch/search?q=&f[ehs_publication_year][0]=**2015**&f[ehs_unit_sort][0]=**2015-10-21**&v=all&group=**bel-244931**&start=0&rows=20&sort=ehs_publication_date%20desc

Cependant, deux limites se sont imposées :

- Comment identifier les cotes des sites archivés ?
- Le nombre de résultats retourné sur la page pour une requête est celui du nombre de pages archivées, et non pas celui du nombre de sites archivés.

En poussant encore un peu plus nos investigations, nous sommes parvenues à identifier l'appel à une API[5] permettant d'interroger les résultats. Mais nous n'avons pas retenu cette méthode, car l'API semble être configurée pour ne donner que 25 résultats par page. Surtout, les informations livrées et leur format étaient trop différents de ce que nous parvenions à récupérer chez IA et empêchaient donc toute comparaison.

3.4.3 Extraction des données par l'équipe technique de la BN

Tant pour Helveticat que pour e-Helvetica Access, les méthodes de collecte explorées se sont avérées trop fastidieuses – ce qui aurait pu être source d'erreur dans la collecte –, mais surtout impossibles à effectuer de manière automatisée.

Nous avons donc assez tôt pris conscience du besoin d'aller chercher l'information à la source, à savoir directement auprès de la BN. Alors que nous avions déjà pris contact dès le mois de mai 2019 avec Madame Barbara Signori, responsable e-Helvetica à la BN, pour l'informer de notre démarche, nous lui avons adressé une requête s'il était possible d'obtenir les métadonnées requises pour nos analyses – si elles existaient – début juillet 2019 (à la suite du constat de la modification de e-Helvetica Access). En raison de circonstances exceptionnelles, notre demande n'a pas pu être prise en charge durant l'été. Nous en avons profité pour détailler plus précisément notre besoin afin que l'institution puisse identifier au mieux les données à nous communiquer. En effet, une extraction de ce type n'était pas vraiment prévue dans le catalogue, elle a donc dû être paramétrée spécifiquement par l'équipe technique de la BN. Un premier exemple de fichier nous a été transmis fin septembre, que nous avons validé. Puis, début novembre 2019, nous avons reçu un second échantillon, composé d'un fichier XML et d'un dossier de fichiers CDX. Tandis que le fichier XML contient à la fois les informations de catalogage en MARC21 et les métadonnées attachées aux sites permettant le suivi de leur moissonnage, les fichiers CDX permettent de prendre connaissance des opérations de moissonnage effectivement réalisées. Ces deux types de données étaient donc complémentaires et répondaient tout à fait à nos souhaits. Une fois un accès FTP ouvert sur le serveur de la HEG Genève, la BN a procédé au dépôt des fichiers entre le 29 novembre et le 4 décembre 2019. À partir de ces données, nous avons pu composer les échantillons BN1 et BN2 (voir rubrique 3.7.1). Le dépôt et le stockage de ces données sur le serveur de la HEG Genève nous a permis de les exploiter directement dans Dataiku.

[5] https://www.e-helvetica.nb.admin.ch/api/search/?q=&t[0]=web-all&v=all&start=0

3.5 Les API d'Internet Archive

Internet Archive s'est toujours présenté comme un acteur ouvert. Il favorise l'utilisation libre de ses ressources, tant pour la collecte de contenus que pour leur exploitation, et encourage la participation de tout développeur à la création d'outils. Un blog d'IA[6] est même dédié à cette communauté et à ses projets, en plus des nombreux dépôts dont le sujet est Internet Archive sur la plateforme Github[7]. C'est la raison pour laquelle il existe pléthore d'outils, qu'ils soient développés par l'institution elle-même ou par des individus externes, en fonction d'un besoin en particulier.

En plus de ces programmes informatiques ouverts, Internet Archive propose également des API permettant le dépôt, la mise à jour ou la consultation de fichiers. En revanche, leur utilisation est très souvent limitée à des contenus autres que les sites web archivés. Cependant, nous en avons identifié et utilisé deux qui correspondaient à nos besoins.

3.5.1 Interrogation du Wayback CDX Server par URL

L'API « Wayback CDX Server API »[8] permet d'interroger les index des données de capture sous format CDX. Elle nous a été utile dans le cadre de deux processus de collecte. Cette API est utilisée par la Wayback Machine pour présenter les résultats de recherche correspondant à un item. Elle est interrogeable à partir d'une simple URL :

```
http://web.archive.org/cdx/search/cdx?url={{URL}}
```

La requête retourne pour chaque variable {{URL}} une partie de l'index CDX de chaque capture disponible dans l'archive. À partir d'un script rédigé par Bastien Berger, assistant à la HEG, et exécuté directement depuis notre outil d'analyse de données Dataiku, nous avons récupéré, pour la liste des sites archivés par la BN, les métadonnées d'IA (échantillon IA2).

3.5.2 Interrogation du Wayback CDX Server avec des paramètres avancés

D'après la documentation disponible, d'autres paramètres peuvent être ajoutés à l'URL de requête de l'API « Wayback CDX Server API » afin de filtrer plus précisément les résultats. Nous avons ainsi construit une requête plus complexe permettant d'identifier l'ensemble des sites en ccTLD .ch archivés par IA, et ainsi constituer notre ensemble de données IA1. La première étape a consisté à consulter l'URL suivante pour déterminer le nombre de pages à interroger :

```
http://web.archive.org/cdx/search/cdx?url=*.ch&collapse=urlkey&filter=urlkey:.
*\)/$&filter=statuscode:200&filter=mimetype:text/html&showNumPages=true
```

Puis, la seconde à consulter chacune des pages afin d'en extraire les données correspondant aux paramètres que nous avions sélectionnés, sous format JSON, avec l'URL suivante :

```
http://web.archive.org/cdx/search/cdx?url=*.ch&output=json&collapse=urlkey&
filter=urlkey:.*\)/$&page=0&filter=statuscode:200&filter=mimetype:text/html&P
age={{incremantalpagenumber}}
```

[6] https://blog.archive.org/developers
[7] https://github.com/search?q=internet+archive
[8] https://github.com/internetarchive/wayback/tree/master/wayback-cdx-server

Les paramètres en question permettaient de ne récupérer que les résultats pour des sites avec pour TLD .ch (variable `url=*.ch`), limités aux noms de domaine de premier niveau en filtrant sur le SURT dans la colonne *urlkey* à l'aide de l'expression régulière : `.*\)/$&p`. Nous n'avons appelé que les contenus de type texte/html, ce format étant le standard pour l'affichage de contenu sur le web. Enfin, comme mentionné dans la section 1.4, il ne s'agissait de récupérer que les informations de moissonnage correspondant à une réponse valide de la part du site web interrogé (statut HTTP 200). L'exécution de ce script s'est effectuée de manière indépendante, puis la compilation des réponses obtenues a été chargée dans Dataiku. Il convient de rappeler ici qu'IA archive tous les contenus trouvés, même si les créateurs des sites s'y opposent à l'aide d'un fichier robot.txt. Mais dans ce cas, les informations ne sont pas accessibles par la Wayback Machine et l'API ne donnera donc pas de réponses pour ces objets non plus.

3.5.3 Recherche par l'URL de requête de la Wayback Machine

Les informations obtenues par la première méthode nous paraissant sommaires, nous avons exploré une autre piste. À partir d'une page de résultats "Summary" de la Wayback Machine, nous avons identifié l'URL d'interrogation d'une autre API, permettant de récupérer des données sur les versions des sites archivées par IA. L'analyse du code de la page a révélé une balise <script> contenant le paramètre suivant :

```
api_url_anchor": "/__wb/search/anchor?q={{query}}
```

En complétant la variable avec l'URL de la Wayback Machine, nous avons composé l'URL de requête suivante :

```
https://web.archive.org/__wb/search/anchor?q={{query}}
```

Cette API n'étant pas documentée, nous avons seulement pu l'exploiter telle quelle ; il aurait été trop fastidieux d'en imaginer toutes les clefs et fonctionnalités. Cette requête, insérée dans un script de notre programme de traitement des données Dataiku, a notamment permis de récupérer l'année de première et de dernière capture, mais également le nombre de fichiers (vidéo, audio ou image) sauvegardés à partir d'une liste de noms de domaines. Cependant, en consultant *a posteriori* les informations recueillies d'un certain nombre de sites, et en les comparant aux données affichées dans la Wayback Machine, nous avons constaté que les données reçues ne correspondaient pas. De plus, les informations supplémentaires collectées n'avaient pas d'utilité avérée pour notre analyse. Nous avons donc préféré ne pas utiliser ce jeu de données pour nos analyses.

Pour résumer, les sources d'informations sur les outils permettant la collecte des données d'IA sont diverses et dotées d'une documentation plus ou moins complète et actualisée. Par conséquent, il a été difficile de déterminer la meilleure manière d'obtenir des données exploitables. Après de nombreux tests et tentatives de paramétrages, nous avons retenu parmi les différentes méthodes explorées celles qui correspondaient le plus aux besoins de l'analyse et à notre capacité à les mettre en pratique.

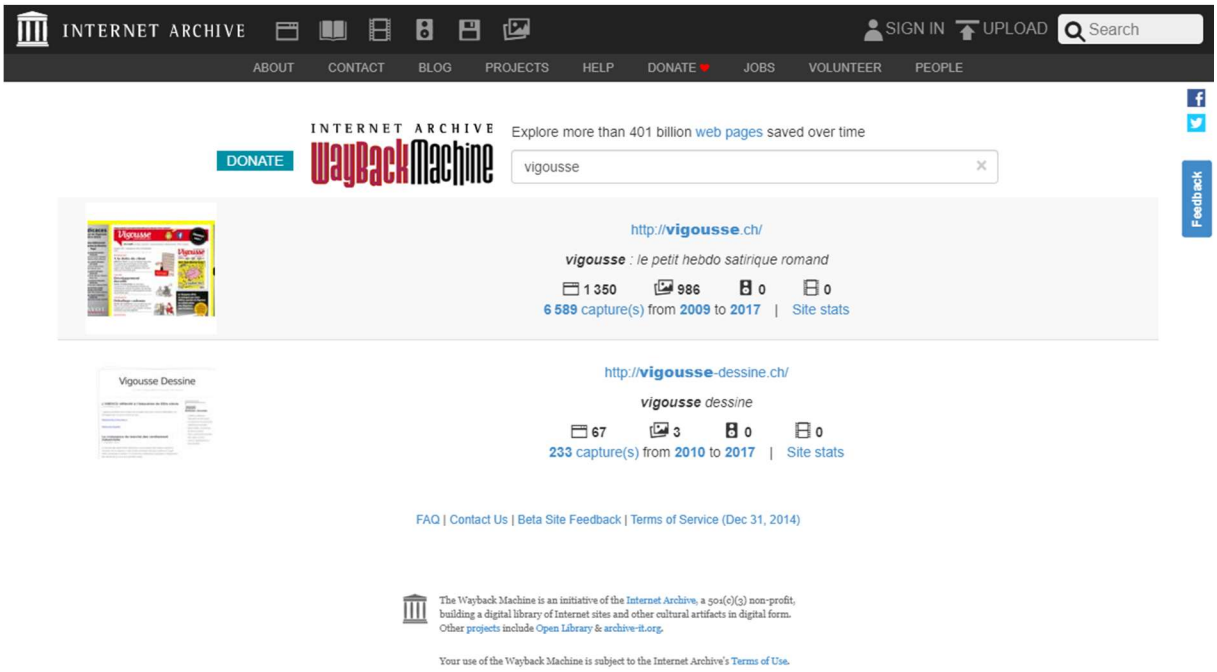
3.6 Consultation des collections et de leurs contenus

Les deux institutions disposent chacune d'une interface web permettant d'explorer leurs archives.

3.6.1 Wayback Machine d'IA

Pour IA, il faut saisir dans la *Wayback Machine* soit une URL exacte, soit un mot qui fait partie de l'URL ou de la description du site.

Figure 6 : Capture d'écran du résultat d'une recherche par mot-clef dans la Wayback Machine

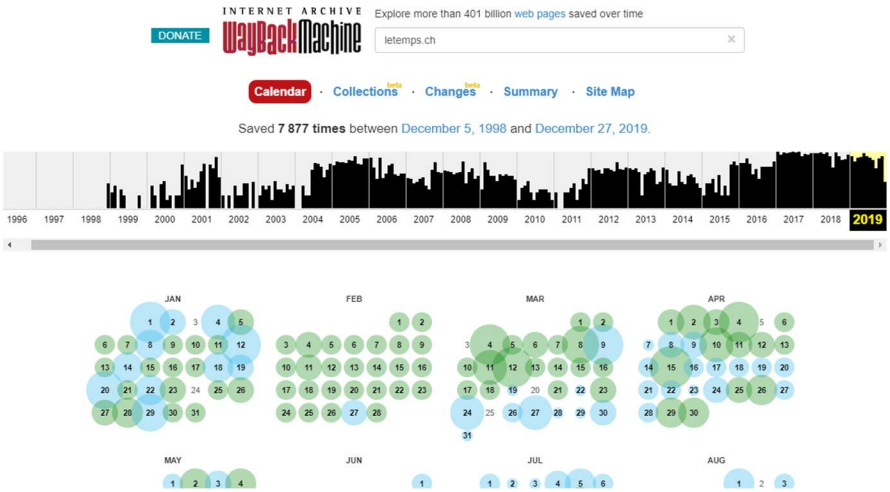


Les résultats pour une URL se présentent sous forme de calendrier, avec des pastilles de couleur en fonction des codes de réponse HTTP sur les dates auxquelles un moissonnage a eu lieu (Internet Archive 2019c) :

- bleu : réponse par le serveur (code 2nn),
- vert : redirection du lien (code 3nn),
- orange : *client error* (code 4nn),
- rouge : *server error* (code 5nn).

Dans la partie supérieure de la page de résultats, une *timeline* indique la répartition des captures dans le temps. On peut cliquer dessus pour sélectionner une version à visionner. Par défaut, les contenus archivés sont consultables en ligne.

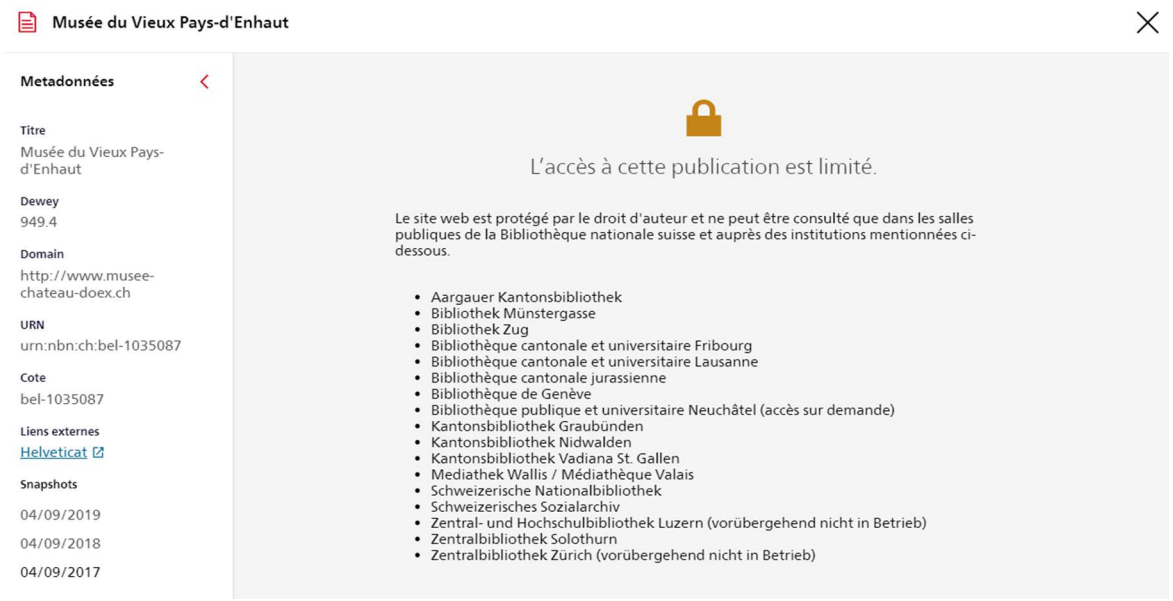
Figure 7 : Capture d'écran de la visualisation des moissonnages existants pour une URL sous forme de *timeline* et calendrier



3.6.2 e-Helvetica de la BN

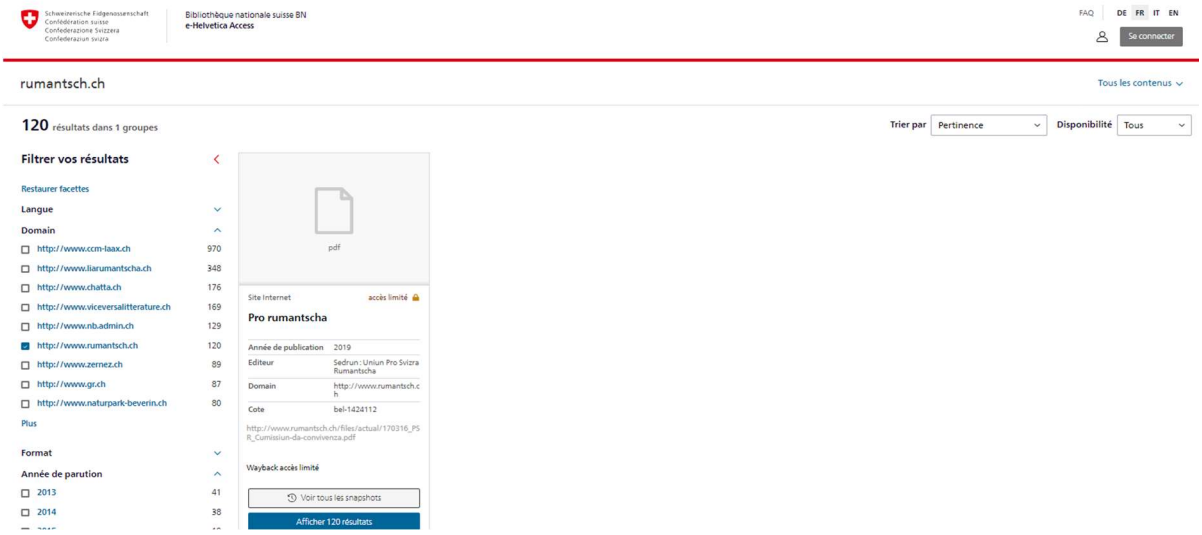
Dans e-Helvetica, l'interface de la BN, seules les métadonnées des sites archivés sont publiquement visibles (voir Figure 8). Pour consulter des contenus il faut se rendre dans une des institutions patrimoniales partenaires.

Figure 8 : Capture d'écran du catalogue en ligne e-Helvetica, indiquant la limitation de l'accès aux ressources



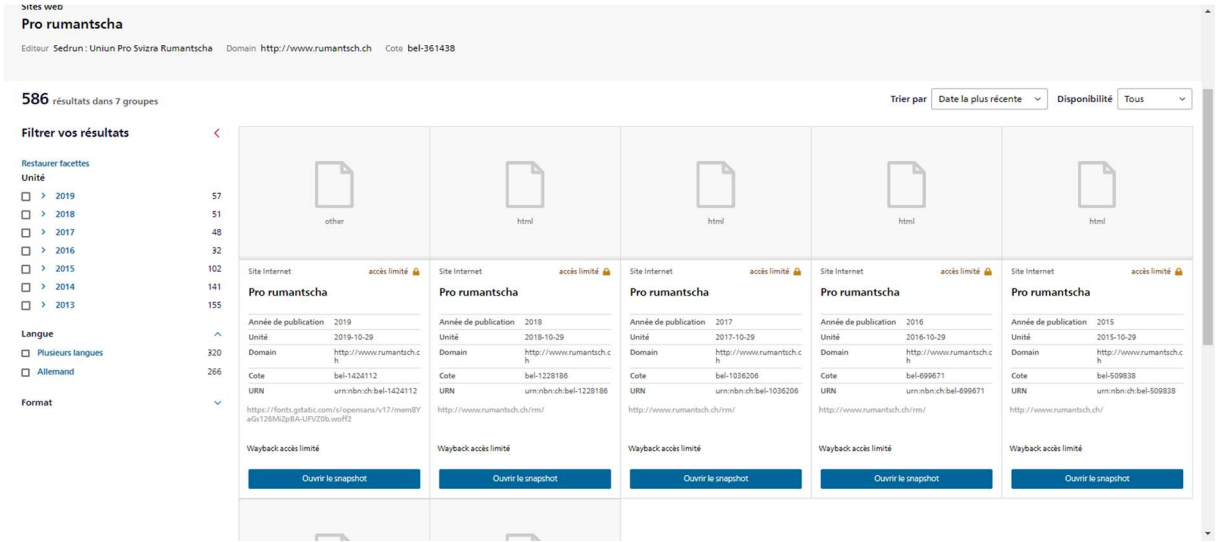
La recherche se fait par URL complète ou par mot-clef, les résultats peuvent ensuite être filtrés par année de publication, par langue, etc. La liste des résultats comporte également des sites qui ont un lien vers l'URL recherchée, c'est pourquoi il convient de cocher celle-ci dans le filtre "Domain".

Figure 9 : Capture d'écran du résultat d'une recherche dans e-Helvetica Access après application du filtre par *Domain*



Un clic sur « Voir tous les snapshots » permet ensuite d'afficher les métadonnées des différentes captures, notamment la date de moissonnage et le nombre de pages archivées (voir Figure 10).

Figure 10 : Capture d'écran de la vue des différents snapshots

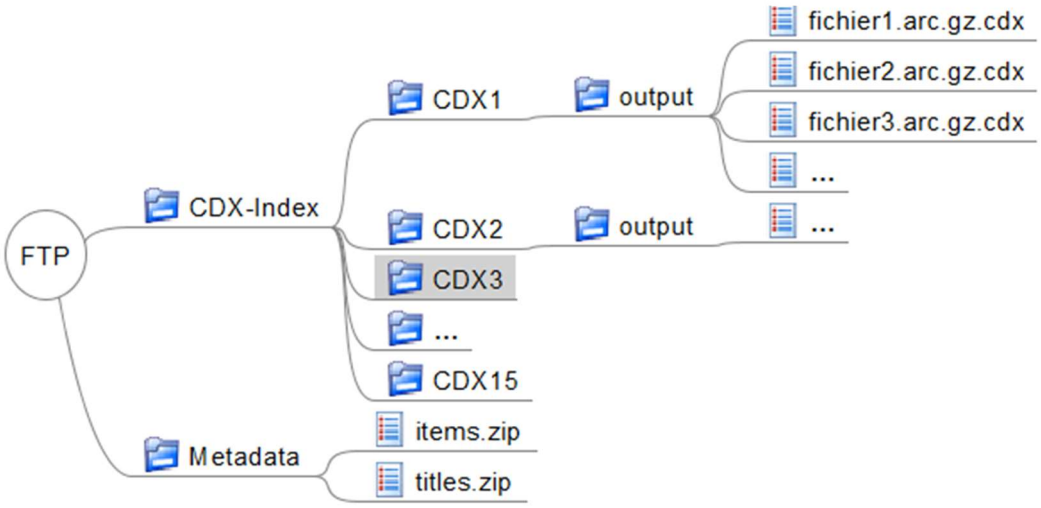


3.7 Description des données recueillies

3.7.1 Les données de la BN

Les données de la BN à analyser nous ont été fournies par l'institution elle-même, par le biais d'un dépôt des fichiers par protocole FTP sur le serveur de la HEG (voir rubrique 3.4.3). L'arborescence se présentait de la manière suivante :

Figure 11 : Arborescence des données livrées par la BN



Les fichiers items.zip et titles.zip du dossier Metadata contenaient des fichiers XML. Ils ont été dézippés directement depuis le serveur afin de pouvoir en importer le contenu dans Dataiku pour analyse et traitement.

Les fichiers XML sont nommés selon la cote en usage dans les catalogues de la BN, à savoir le préfixe « bel- » suivi d'un identifiant unique. Ils contiennent d'une part les informations de catalogage au format MARC21, et d'autre part les informations techniques relatives au moissonnage des sites pour archivage. Alors que les fichiers « titles » traitent des sites en eux-mêmes, les fichiers « items » portent sur les instances de capture. Ils renferment des renvois vers les autres versions archivées du même titre, à l'aide de leurs identifiants (qui est équivalent à leur URN et leur cote) et de liens pérennes enregistrés auprès d'un résolveur de liens.

Parmi les informations contenues dans les balises XML, on retrouve notamment :

- les cotes des sites et de leurs moissonnages (identifiants uniques),
- le titre donné par la BN aux sites,
- les URL sondées,
- les langues des sites,
- l'identification de l'institution qui a soumis la proposition de moissonnage,
- une date de *harvest,*
- les liens entre les différentes captures et le titre d'un site.

Certaines de ces informations n'ont pas d'équivalent dans les données collectées auprès d'IA et ne peuvent donc être exploitées dans le cadre de notre analyse quantitative, qui porte sur la comparaison entre ces deux acteurs de l'archivage du web. Néanmoins, il est intéressant de les mentionner pour un éventuel autre projet de recherche plus spécifique à la BN.

Par ailleurs, nous avons également identifié que la "date de moissonnage" incluse dans les données techniques n'était en réalité que celle de l'enregistrement de l'annonce de site, et non pas celle à laquelle le premier moissonnage a effectivement été réalisé. Cela est flagrant pour les sites déclarés en 2008, au lancement du projet Archives Web Suisse et capturés en 2009. Il serait donc préférable d'utiliser les informations de capture effective contenues dans les fichiers CDX.

Pour ce qui est des dossiers CDX, chacun contient exactement un dossier *output*, qui comporte, lui, un nombre variable de fichiers avec extension .arc.gz.cdx. Le dossier intermédiaire n'a pas d'utilité et doit résulter d'un traitement effectué par la BN pour rassembler les données. Nous ne le mentionnerons donc plus dans la suite de ce document. Le format de fichier CDX permet de décrire des documents web archivés. La première ligne du CDX fournit la légende, sous forme codée (voir Tableau 2), des données descriptives contenues dans les lignes suivantes.

Tableau 2 : Légende des codes des champs CDX des fichiers de la BN

Code	Correspondance	Définition
N	massaged url	Le document capturé, désigné par son URL convertie en SURT
b	date	Date et heure à laquelle le document a été capturé
a	original url	Le document capturé, désigné sous son URL originale
m	mime type of original document	Le type de fichier du document
s	response code	Le code de réponse HTTP du document lors de la capture
k	new style checksum	La valeur de somme de contrôle SHA-1 unique codée en Base32 pour le document, pour le distinguer des autres (alternative au MD5)
r	redirect	-
M	meta tags (AIF)	-
S	compressed record size	La taille du document en octets dans sa forme d'enregistrement compressé
V	compressed arc file offset	La taille du document en octets dans son fichier d'archive web WARC
g	file name	Le nom du fichier d'index contenant les informations

(IIPC 2015 ; Blumenthal 2018)

Il y a 15 dossiers CDX. Dans un premier temps, nous avions pensé que chacun correspondait à une année particulière de moissonnage, mais une brève réflexion nous a permis de réaliser que ce n'était pas possible : le nombre de dossiers ne correspondait pas au nombre d'années depuis lesquelles la BN procède à l'archivage de sites web. De plus, le nommage des fichiers comportant parfois le nom du site archivé et une date, nous avons remarqué que certains fichiers apparaissaient dans plusieurs des dossiers. Nous avons donc appliqué par la suite des opérations de dédoublonnage des données.

Une particularité, que nous n'avons constatée malheureusement que dans un second temps, est une grande différence numéraire entre le nombre de résultats du catalogue en ligne Helveticat (12'180) et le nombre de fichiers XML de type Title (12'072) d'une part, le nombre d'URL contenues dans les fichiers CDX (30'197) d'autre part. Il s'avère que les CDX de la BN contiennent davantage d'URL que celles réellement archivées par cette institution. Renseignement pris auprès de Mme Signori, le crawler, en suivant les liens contenus dans un des sites programmés pour moissonnage, va, lorsqu'il tombe sur un lien externe, « jeter un coup d'œil dehors ». Cette mini-visite d'une page relevant d'un autre domaine donne lieu à la création d'un fichier CDX, mais l'URL externe en question n'est évidemment pas cataloguée dans Helveticat, ni accessible *via* l'interface des Archives Web Suisse. Si l'on se réfère à la norme ISO 14873 (Organisation Internationale de Normalisation 2013), le catalogage des sites archivés est un exercice optionnel dans l'absolu. Cependant, l'intention de la BN est de signaler toutes ses ressources en les cataloguant (même si les contenus eux-mêmes sont inaccessibles), intention corroborée par sa pratique de collecte sélective et le formulaire d'annonce de site à archiver, qui contient des champs similaires aux notices de catalogue. Ainsi, il convient de retenir seulement les URL cataloguées lors de la préparation des données.

3.7.2 Les données d'Internet Archive

Contrairement à la BN dont nous avions l'entier des données, pour IA, nous avons renouvelé nos requêtes pour chaque corpus. En effet, en raison du très grand volume de sites archivés par IA, mais aussi du fait que l'archivage est un processus continu et automatisé, il aurait été impossible d'obtenir la totalité des données. Nous avons donc procédé à une collecte à chaque moment où le besoin s'en faisait ressentir.

Les données recueillies auprès d'Internet Archive sont celles des fichiers CDX générées à chaque opération d'archivage d'un contenu. Les informations ont été collectées grâce à une API spécifique auprès du serveur d'IA dédié à l'interrogation des fichiers CDX, et dont la configuration délivre les métadonnées de collecte structurées suivantes (voir Tableau 3) (Kreymer et al. 2018) :

Tableau 3 : Détail du contenu des fichiers CDX d'IA

Attribut	**Définition**
urlkey	Le document capturé, désigné par son URL convertie en SURT
timestamp	Date et heure à laquelle le document a été capturé
original	Le document capturé, désigné sous son URL originale
mimetype	Le type de fichier du document
status code	Le code de réponse HTTP du document lors de la capture

digest	La valeur de somme de contrôle SHA-1 unique codée en Base32 pour le document, pour le distinguer des autres (alternative au MD5)
length	La taille du document en octets dans son fichier d'archive web WARC

(Blumenthal 2018)

Certaines colonnes des fichiers CDX correspondent entre la BN et IA (massaged URL et urlkey, date et timestamp, original URL et original) permettant leur comparaison. De manière générale, nous avons utilisé comme identifiant unique des sites leur SURT, ou leur nom de domaine, l'un étant créé à partir de l'autre.

3.8 Préparation des données

À plusieurs reprises, pour nous assurer de comparer ce qui est comparable, nous avons procédé à des opérations de sélection et de formatage des données.

3.8.1 Traitement des données CDX

Pour les échantillons CDX des deux institutions, quelques opérations de normalisation facilitant l'analyse des données ont été appliquées dans Dataiku, notamment pour extraire le nombre de nouveaux sites moissonnés par année.

Les étapes de préparation ont consisté à :

1. Parser le champ de date et heure de collecte pour extraire l'année
2. Parser les URL d'origine pour en extraire le nom de domaine original (et permettre ultérieurement la comparaison entre les deux sets), mais parfois aussi identifier les différents ccTLD
3. Supprimer les données à double
4. Rassembler les lignes selon leur URL originale et les ordonner par année croissante, afin de ne conserver, pour chaque URL, que la première opération de moissonnage.

Ces opérations ont été appliquées aux ensembles de données IA1 et IA2, mais aussi aux données CDX de la BN ayant permis d'établir les ensembles BN1 et BN2.

3.8.2 Création des échantillons BN1 et BN2

L'ensemble de données obtenues auprès de la BN contient les informations de capture de l'ensemble des sites qu'elle archive et catalogue. Si l'ensemble BN1 comprend les informations de ceux correspondant à nos critères de filtrage, tous TLD confondus, l'ensemble BN2 en est une extraction ne contenant que les sites au ccTLD .ch.

La constitution des ensembles de données à analyser s'est donc déroulée en plusieurs étapes :

1. Importation des données issues des 15 dossiers de fichiers CDX dans Dataiku et nettoyage afin de ne conserver que les informations répondant aux mêmes critères que ceux utilisés lors de l'interrogation d'IA, à savoir :
    a. noms de domaines de premier niveau (filtre sur la colonne du SURT avec l'expression régulière `^[a-zA-Z]+,[\w-]*\)\/$`),
    b. statut de réponse HTTP valide (200),

c. type de fichier text/html,

d. informations de la première capture uniquement.

2. En recourant au programme dnGREP et à l'expression régulière <nlnz:Process>http.*<\/nlnz:Process>, interrogation des fichiers XML contenus dans le dossier Metadata\Titles pour en extraire les balises <Process> et leur contenu, précédemment identifiées comme renfermant à la fois les URL des sites moissonnés, et les URL pérennes des ressources. Importation du résultat de cette extraction dans Dataiku. Nettoyage en supprimant les balises et les lignes correspondant aux liens pérennes (la base de l'URL étant identique à toutes les URL).

   C'est ainsi que nous sommes parvenues à identifier la liste complète des sites archivés par la BN (selon nos critères), soit 10'955. Nous avons ensuite croisé cette liste aux données CDX de la BN préparées tel que décrit dans la rubrique 3.8.1 pour ne garder que les informations de capture des "Helvetica" et constituer ainsi l'ensemble de données BN1, soit 8'132 URL. Il est à préciser que cette opération conduit à la perte de 2'823 références (environ un quart). Cette perte est liée au fait que dans les fichiers de catalogage XML, les balises retenues pour l'extraction ne mentionnent que le domaine de premier niveau, alors que la capture effective renseignée dans le CDX porte sur des URL plus grandes.

3. Pour l'analyse propre au ccTLD .ch, isolement à partir de l'ensemble BN1 des lignes pour lesquelles le domaine appartient effectivement au ccTLD en question. Le résultat est l'échantillon BN2.

3.9 Objet de l'analyse

Il existe plusieurs manières d'évaluer une collection d'archives du web : en analysant les actions d'archivage d'une institution ou en la comparant à d'autres institutions, en appliquant une analyse qualitative ou une analyse quantitative.

En fonction de l'objectif de notre travail de recherche (voir sections 1.2 et 1.3) et des données que nous étions capables de collecter, et tout en nous référant à la norme ISO 14873, nous avons sélectionné nos propres critères d'analyse selon la nature de notre étude.

Ainsi, pour l'analyse quantitative, nous avons évalué :

- l'identification, basée sur les domaines de premier niveau moissonnés,
- la présence, garantie dans les informations de capture par la réponse HTTP 200 aux requêtes mais également par la date de capture la plus ancienne extraites des données CDX,
- la représentativité, en comparant le volume de sites archivés par rapport au nombre existant, ou encore en comparant les collections de la BN avec celles d'IA.

Pour des raisons pratiques, mais également en raison du volume de données trop important, nous avons préféré ne pas explorer les questions de fréquence, de complétude et de profondeur des objets archivés dans notre analyse quantitative, mais seulement dans celle qualitative portant sur un échantillon de sites (voir chapitre 5). Nous avons écarté de notre analyse certains autres critères, tels que les sujets ou la langue des sites, ou encore l'identification des initiateurs d'archivage. Ces informations figurent en effet dans les fichiers XML fournis par la BN, mais n'ont pu être collectées auprès d'IA.

3.10 Évaluation des biais

Comme toute recherche, celle-ci est sujette à des biais qui peuvent altérer la justesse de nos résultats. Ceux que nous avons identifiés sont :

- une potentielle falsification involontaire dans le traitement des données ; en effet nous sommes novices dans le traitement des données, et nos choix de filtrage peuvent avoir conduit à des imprécisions dans la sélection des données à analyser. Cela nous est d'ailleurs arrivé lors de l'élaboration du poster de recherche (voir 3.7.1 et 4.1) ;
- la sélection ainsi que la quantité de sites pour l'analyse qualitative ont été définies d'une part à partir de notre expérience personnelle et professionnelle, et d'autre part pour des critères de faisabilité. Cette sélection ne saurait donc être représentative du web suisse réel ;
- les données fournies pour le nombre de sites du web dans sa globalité proviennent d'un site dont nous n'avons pas pu éprouver la fiabilité.

4. Résultats quantitatifs

Pour l'angle quantitatif de notre recherche, nous avons sondé et apprécié le volume de sites web suisses (selon les définitions établies en section 3.1), afin d'évaluer dans quelle proportion leur archivage était assuré par la BN et IA. Nous avons aussi comparé les sites archivés par les deux institutions dans les deux sens : ce qui est archivé par X l'est-il aussi par Y, et vice-versa.

Ces comparaisons portent surtout sur la présence des sites dans les archives des deux institutions, mais analysent aussi les variations de l'accroissement par l'archivage de nouveaux sites à travers le temps, sans entrer dans des questions de performance.

4.1 Comparaison IA vs. BN

Cette analyse a été conduite une première fois début décembre afin d'obtenir des chiffres et un graphique, que nous avons utilisés pour le poster présenté le 12 décembre 2019 (voir Annexe 1). Or, malheureusement, nous avons réalisé par la suite que les données exploitées ne correspondaient pas à ce que nous voulions étudier (voir rubrique 3.7.1). En effet, la pression due aux difficultés techniques et au manque de temps avant la livraison du fichier pour impression a altéré notre raisonnement.

Ainsi, nos résultats obtenus en comparant l'ensemble des URL des CDX de la BN avec ceux d'IA n'étaient pas corrects. Nous nous sommes résolues à réitérer la préparation des données de la BN (voir rubrique 3.8.1). Pour cette analyse, nous avons donc comparé les ensembles de données IA1 et BN2.

4.1.1 Situation à l'instant t

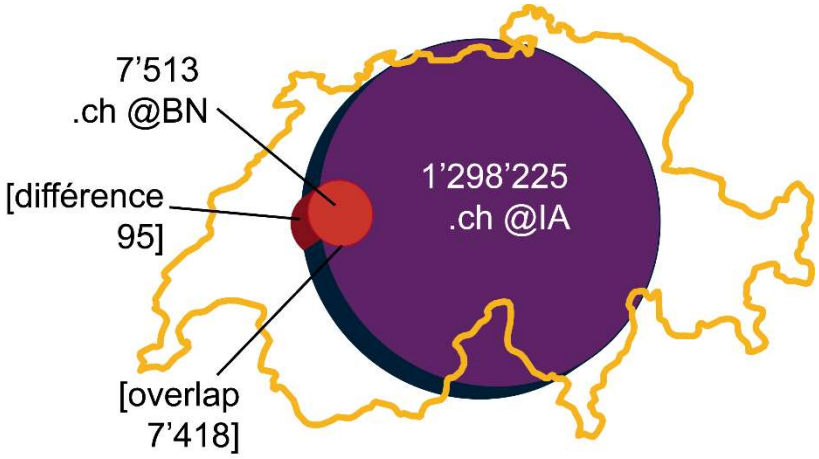
Le traitement des données nous a permis d'établir qu'à l'automne 2019, le nombre de sites de ccTLD .ch archivés par IA était de 1'298'225, tandis que la BN en comptait 7'513.

En proportion, le volume des sites .ch archivés par la BN ne représente que 0.58 % du volume des sites archivés par IA. Si l'on rapporte ces volumes au nombre total de sites .ch enregistrés[9] au 15 décembre 2019, qui s'élève à 2'259'952, IA en archive 57.44 %, ce qui constitue un score raisonnable, tandis que la BN atteint uniquement 0.33 %.

En croisant les URL archivées par IA d'une part et par la BN d'autre part, il s'avère que 7'418 sites sont archivés par les deux institutions (voir Figure 12). Ce qui signifie que, malgré le nombre plus réduit de sites traités par la BN, elle en compte néanmoins 95 qui ne sont pas archivés par IA.

[9] https://www.nic.ch/fr/statistics/domains/

Figure 12 : Proportion et recouvrement de sites .ch archivés selon les institutions



En analysant ces 95 sites, on constate que la majorité des sites manquants chez IA ont été capturés pour la première fois par la BN tout récemment (voir Tableau 4). Une hypothèse pourrait donc être qu'il s'agit de sites récents et que les *crawlers* d'IA n'ont pas encore eu l'occasion de passer. Une autre raison pourrait être que les sites en question sont peu reliés au reste du web par des liens entrants, et que les *crawlers* ne peuvent donc les atteindre. Ces hypothèses restent à vérifier dans un autre projet de recherche.

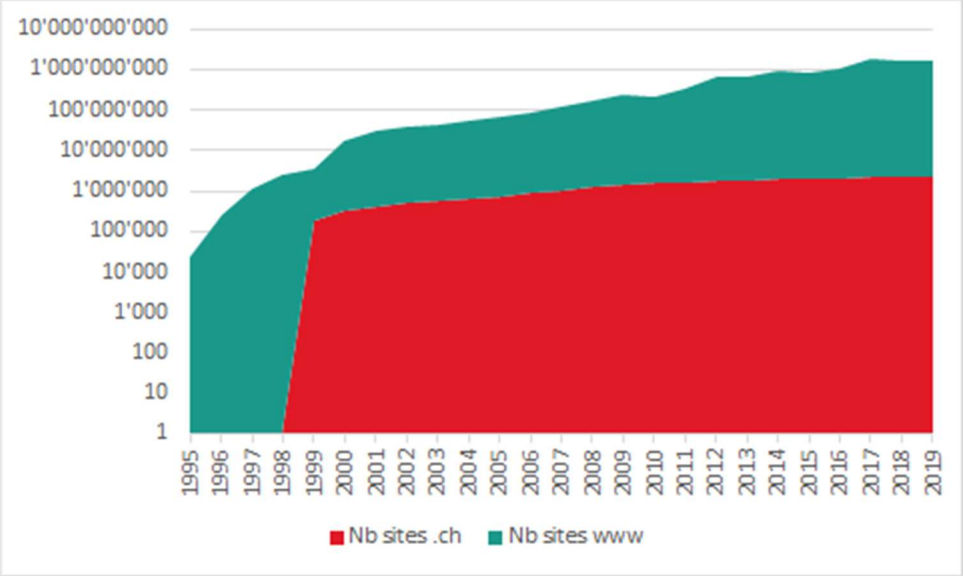
Tableau 4 : Répartition des sites non-archivés par IA en fonction de l'année de première capture par la BN

Année de collecte à la BN	**Nombre de sites uniquement archivés par IA**	**Proportion**
2019	46	48.4 %
2018	15	15.8 %
2017	13	13.7 %
2016	7	7.4 %
2015	2	2.1 %
2014	2	2.1 %
2013	7	7.4 %
2012	1	1.1 %
2011	2	2.1 %

4.1.2 Le World Wide Web et le ccTLD .ch

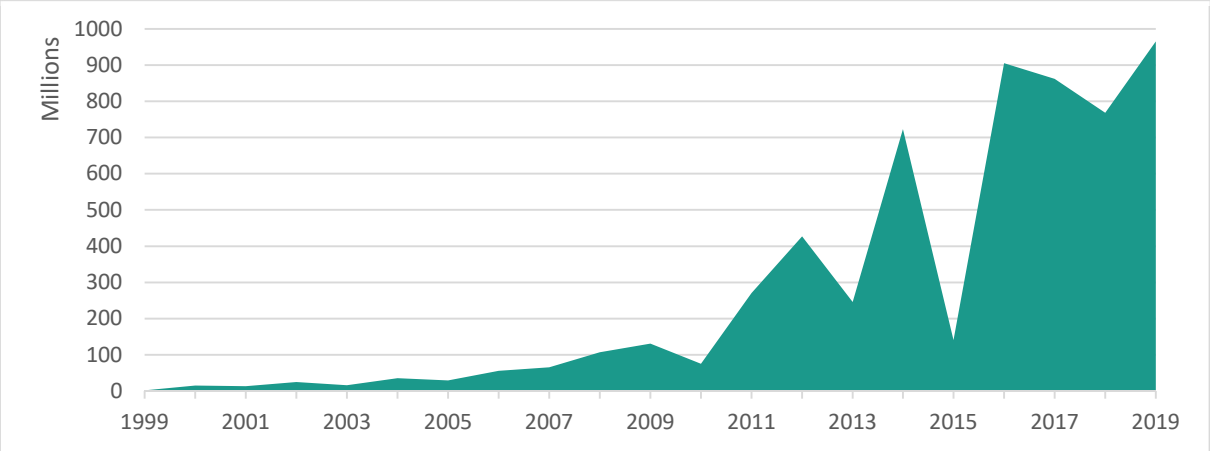
La croyance populaire veut que le volume du web augmente de manière exponentielle à travers le temps. Lorsqu'on observe la courbe de croissance du web depuis sa création (voir Figure 13), ce constat est valable. Il en est de même si l'on considère le nombre de sites au ccTLD .ch enregistrés auprès de Switch.

Figure 13 : Nombre de sites .ch vs. l'ensemble du web par année (volume en échelle logarithmique)



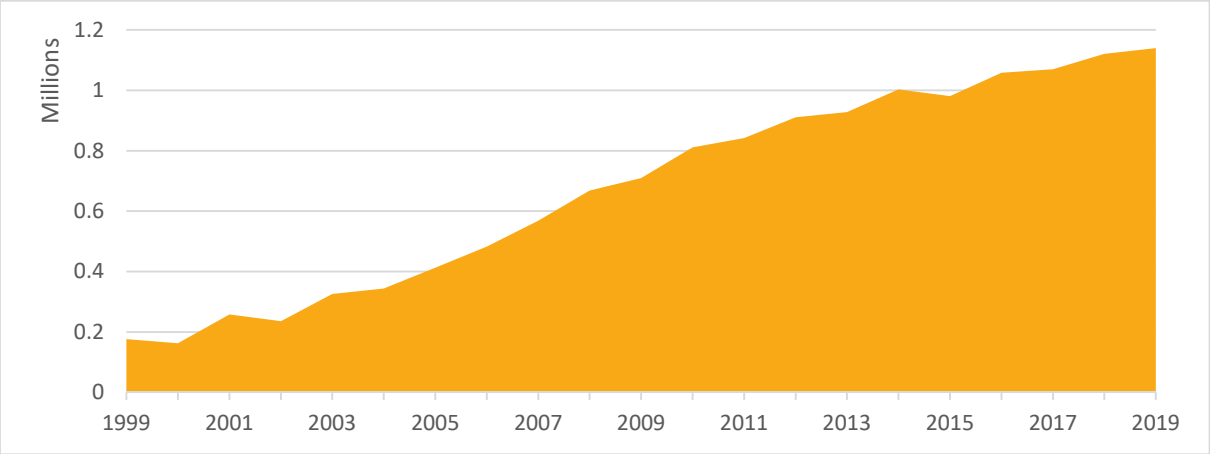
Cependant, en regardant plus près, il convient de souligner que ces augmentations ne se produisent pas de manière régulière, mais par à-coup. Ce constat s'applique tant pour le web global (voir Figure 14) qu'au niveau du ccTLD .ch (voir Figure 15). Dans ce deuxième cas, les fluctuations sont moins brutales, mais existent bel et bien, selon les statistiques collectées auprès du *registrar* Switch (chiffres en Annexe 2).

Figure 14 : Courbe d'évolution du nombre de nouveaux sites du World Wide Web par année



Adaptation de https://www.internetlivestats.com/total-number-of-websites/#trend

Figure 15 : Courbe d'évolution du nombre de nouveaux sites au ccTLD .ch enregistrés par Switch par année

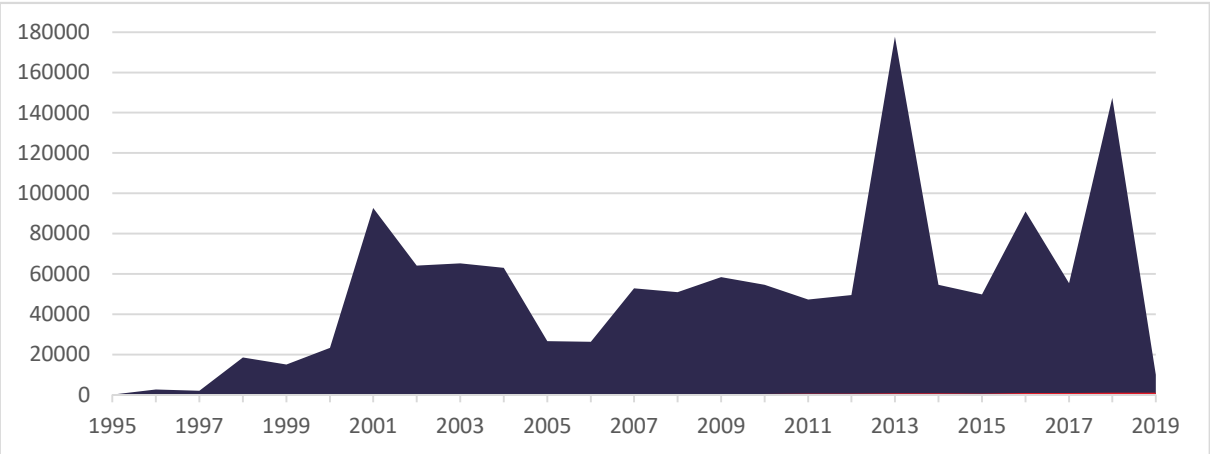


Adaptation de https://www.nic.ch/fr/statistics/domains

4.1.3 Évolution à travers le temps

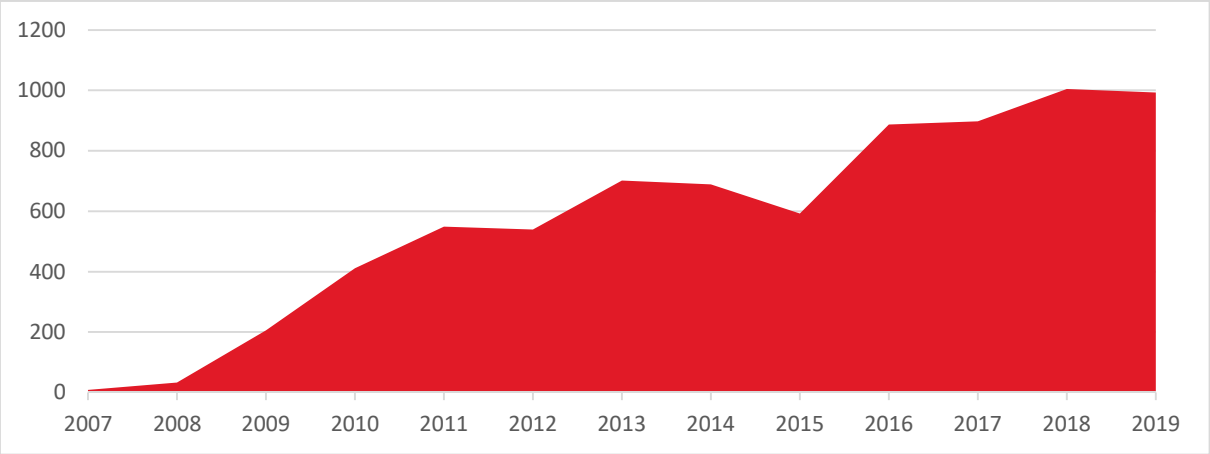
Chez IA, le processus étant entièrement automatisé, on pourrait légitimement imaginer que le nombre de sites nouvellement archivés suit une courbe régulière. Mais comme le montre le graphique ci-dessous (voir Figure 16), l'archivage de nouveaux sites ne suit pas la même tendance : on observe des pics et des creux, encore plus prononcés que pour l'augmentation du volume du web global.

Figure 16 : Nombre de nouveaux sites .ch archivés par IA chaque année



À la BN, cependant, les fluctuations sont moins manifestes et la croissance plus régulière (voir Figure 17), à part un petit recul vers l'année 2015.

Figure 17 : Nombre de nouveaux sites .ch archivés par la BN chaque année



Nous n'avons pas d'explication concrète pour ces pics et supposons qu'ils sont dus au hasard, au moins du côté d'IA. Les crawlers de cette institution se promènent à travers le web de manière assez aléatoire. Selon les liens qu'ils suivent, ils « restent en Suisse » – et moissonnent donc du contenu avec le ccTLD .ch – plus ou moins longtemps. Quant à la BN, il est possible que les ressources (humaines et financières) consacrées à l'archivage du web varient d'une année à l'autre, tant à Berne que dans les institutions partenaires qui proposent des sites.

Autre observation intéressante : en 2008, année de l'initialisation de la capture des sites web "Helvetica" par la BN, IA comptait déjà dans ses archives 4'633 des sites .ch de l'actuelle collection de la BN, soit près de la moitié de celle-ci. L'institution suisse a donc bien rattrapé son retard dans la douzaine d'années depuis cette date.

Il est à noter que les 9 sites qui ont servi à la mise en place de l'archivage à la BN en 2007 figurent tous parmi les sites qui avaient déjà été archivés par IA auparavant.

4.2 Comparaison BN vs. IA

Pour cette comparaison, nous avons cherché à évaluer la performance d'IA pour l'archivage de ce que la BN considère comme intéressant à préserver. Nous avons extrait la liste de tous les sites figurant dans les Archives Web Suisse de la BN (tous TLD confondus / ensemble BN1) et regardé s'ils étaient également présents chez IA (ensemble IA2).

4.2.1 État de la collection à la BN

Avant d'entrer dans le vif de cette comparaison, résumons l'état de la collection de la BN en quelques chiffres.

En premier lieu, si le catalogue en ligne Helveticat de la BN fait état de 12'180 notices de sites web archivés, nous n'avons reçu des métadonnées XML que pour 12'072 d'entre eux. Nous imaginons que la différence provient de la complétion des archivages entre le moment où les métadonnées ont été extraites pour nous être livrées et le moment où nous avons consulté le catalogue en ligne.

Parmi ces URL, le nombre de celles qui correspondaient à notre objet d'analyse – à savoir les domaines de premier niveau – s'élevait à 10'955 sites. En revanche, en les croisant avec les

données CDX, nous n'avions d'information de capture que pour 8'132 d'entre eux en tant que nom de domaines principaux (voir rubrique 3.8.2). Ces sites sont enregistrés sous 28 TLD différents (voir Annexe 3), la grande majorité étant des .ch (92,6 %, soit 7'531). Le TLD le deuxième plus courant est l'extension commerciale .com, suivi de .org.

4.2.2 Comparaison

Dans les 8'132 sites archivés par la BN, 8'048 sont également archivés chez IA. Seuls 84 sites sont donc absents, soit environ 1 %.

Le volume total de .ch archivés par la BN étant de 7'531, soit une différence de seulement 1,5% (113 sites) par rapport à ce que nous avons analysé précédemment, nous n'avons pas reconduit toutes les analyses réalisées dans l'analyse précédente (voir rubrique 4.1.1). Cela aurait été redondant et n'aurait pas apporté d'information supplémentaire.

4.3 Analyse

Il ressort de ces deux analyses qu'IA, malgré sa force d'action et l'automatisation de ses processus, n'atteint qu'un peu plus de la moitié du ccTLD .ch. Nous n'avons pas trouvé de statistiques exactes quant au nombre de sites totaux que l'organisation archive effectivement (elle fournit ses chiffres en nombre de pages web). Le ccTLD .ch ne représente qu'une toute petite portion du web. Qu'en est-il des autres TLD ? Les ambitions d'IA manquent-elles de ressources ?

Face au score d'IA, le volume de sites archivés par la BN est maigre. Mais pour rappel, son archivage est le fruit d'une sélection réalisée par des humains répartis en différentes institutions, en fonction d'une politique d'acquisition. La croissance régulière de sites nouvellement collectés peut témoigner que les processus mis en place par la BN sont efficaces, malgré des moyens limités en comparaison avec IA.

Le taux de couverture commune des deux institutions est quasiment 100 %. Mais les données analysées ne concernent que les informations d'une seule capture de chacun des sites, celle de leur premier moissonnage. Afin de juger de la qualité et de la représentativité des archives, il conviendrait d'étudier la fréquence, le volume et la profondeur pour chacun des sites. Pour des raisons de faisabilité, nous ne pouvions mener ces analyses sur la totalité des sites présents. Le chapitre suivant traite donc de l'analyse qualitative sur un échantillon de sites seulement.

5. Résultats qualitatifs

Nous avons dressé une liste d'une vingtaine de sites web[10] (voir Tableau 5) qui, tout en présentant des caractéristiques variées, correspondent à la définition des "Helvetica" selon la loi et l'ordonnance sur la BN[11] : ils traitent de sujets suisses et/ou sont élaborés en Suisse et/ou par des suisses. Notre sélection se veut pluraliste ; il y a des sites d'institutions publiques ou d'ONG, de médias ou de particuliers, à vocation commerciale ou pas, avec différents TLD, dans différentes langues nationales ou multilingues, existant depuis longtemps ou liés à un événement particulier…

Tableau 5 : Liste descriptive des sites examinés pour l'analyse qualitative

#	URL	Description	Langue(s)
1	als.wikipedia.org/wiki/Portal:Basel	encyclopédie collaborative - sous-site dédié à Bâle de la version *alemannisch* de la page Wikipédia	site en suisse allemand
2	carouge.ch	administration d'une commune genevoise	site en français
3	coccinelle.ch	association autour des modèles de collection VW	site en français
4	cosadoca.ch	consortium de sauvetage du patrimoine documentaire en cas de catastrophe, réunissant les Archives cantonales vaudoises, la Bibliothèque de l'EPFL et la Bibliothèque cantonale et universitaire de Lausanne	site en français
5	etoile-filante.biz	site privé - journal d'un voyage en voilier par un couple neuchâtelois	site en français
6	eu-diktat-nein.ch	site de campagne pour la votation fédérale du 19 mai 2019 à propos de la Loi sur les armes	site multilingue de/fr/it
7	heidiland.com	office de tourisme de la région autour du lac Walen (cantons de Saint-Gall et des Grisons)	site multilingue de/en/fr/it
8	helvetas.org	ONG active dans l'aide au développement	site multilingue de/en/es/fr/it
9	hesge.ch	établissement de formation - Haute école spécialisée de Genève	site en français
10	hug-ge.ch	institution parapublique - Hôpitaux universitaires de Genève	site bilingue en/fr

[10] Cette sélection a été établie avant le début de l'analyse quantitative et n'est donc pas influencée par les autres résultats.

[11] Pour rappel, la BN applique des règles plus restrictives à la sélection des sites web qu'aux autres types de documents, cf. Signori (2019b).

#	URL	Description	Langue(s)
11	infoalternativecpeg.org	site de campagne pour la votation cantonale genevoise du 19 mai 2019 à propos de la récapitalisation de la Caisse de pension de la fonction publique à Genève (n'est plus en ligne à fin 2019)	site en français
12	jean-monnet.ch	fondation Jean Monnet pour l'Europe	site bilingue en/fr
13	letemps.ch	média - quotidien suisse romand édité à Genève	site en français
14	lightpainter.ch	portfolio et offre de service d'un photographe	site en français
15	melazic.com	site de vente en ligne (pâtisseries, ateliers créatifs, articles de pâtisserie et objets cadeaux)	site en français
16	nlb.ch	entreprise privée - compagnie de navigation du lac des Brenets	site multilingue de/en/fr
17	rumantsch.ch	association œuvrant pour la langue romanche	site bilingue de/rm
18	strickcafe.ch	site de vente en ligne (laine, accessoires, patrons… pour le tricot), contient également un blog	site bilingue de/fr
19	swiss.com	entreprise privée - compagnie d'aviation Swiss	site multilingue de/el/en/es/fr/it/ja/pt/ru/zh
20	swissair.com	entreprise privée ayant cessé d'exister - ancienne compagnie d'aviation Swissair (domaine repris par Swiss)	site multilingue de/en/fr (selon les périodes)
21	switch.ch	fondation qui gère le réseau académique Suisse ; *registrar* pour les TLD .ch et .li	site multilingue de/en/fr/it
22	vigousse.ch	média - journal satirique hebdomadaire suisse romand	site en français
23	vsa-aas.ch	association professionnelle des archivistes suisses	site multilingue de/fr/it

Nous avons ensuite collecté, à l'aide des interfaces mises à disposition, les éléments suivants pour notre analyse qualitative (voir Tableau 6) :

- présence du site dans la collection (oui / non),
- nombre de versions présentes,
- dates extrêmes (première et dernière capture).

Le nombre de versions présentes a été relevé comme suit :

- pour la BN : le nombre de moissonnages correspond au nombre de dates de capture indiquées sur la page de résultat de e-Helvetica pour un TLD donné ; le total de pages

correspond au nombre de « résultats » indiqué sur cette même page (équivalent à l'addition des pages relatives à chacune des différentes dates de captures) (cf. figure 10 sur la page 23).

- pour IA : le nombre de moissonnages correspond au nombre de dates de capture indiquées sur la page de résultat de la Wayback Machine (« Saved x times between… ») (cf. figure 7 sur la page 22).

Les dates extrêmes des captures ont également été relevées sur les pages de résultat de e-Helvetica et de la Wayback Machine respectivement.

Tableau 6 : Présence des sites à la BN et chez IA

Dernière interrogation : 28.12.2019		BN				IA		
		Nb de versions présentes		Dates extrêmes		Nb de versions présentes	Dates extrêmes	
#	Site	Nombre de moissonnages	Total de pages	Première capture	Dernière capture	Nombre de moissonnages	Première capture	Dernière capture
1	als.wikipedia.org/wiki/Portal:Basel	-	-	-	-	2	07.07.2017	01.12.2018
2	carouge.ch	9	240’348	2010	2017	252	05.12.1998	26.12.2019
3	coccinelle.ch	-	-	-	-	195	05.12.1998	25.05.2019
4	cosadoca.ch	-	-	-	-	57	15.06.2006	03.06.2019
5	etoile-filante.biz	-	-	-	-	18	10.05.2013	09.05.2019
6	eu-diktat-nein.ch	2	658	13.05.2019	24.05.2019	53	01.11.2018	23.11.2019
7	heidiland.com	1	18’874	2019	-	252	13.07.1997	04.10.2019
8	helvetas.org	1	97’378	2019	-	487	21.01.2003	30.11.2019
9	hesge.ch	3	17’511	2017	2019	351	04.02.1998	02.06.2019
10	hug-ge.ch	7	94’158	2011	2019	700	02.10.1999	22.12.2019
11	infoalternativecpeg.org	-	-	-	-	-	-	-
12	jean-monnet.ch	-	-	-	-	162	18.08.2000	10.12.2019
13	letemps.ch	11	612’679	2011	2019	7’877	05.12.1998	27.12.2019

Dernière interrogation : 28.12.2019		BN				IA		
		Nb de versions présentes		Dates extrêmes		Nb de versions présentes	Dates extrêmes	
#	Site	Nombre de moissonnages	Total de pages	Première capture	Dernière capture	Nombre de moissonnages	Première capture	Dernière capture
14	lightpainter.ch	-	-	-	-	16	09.03.2009	08.06.2019
15	melazic.com	-	-	-	-	201	10.08.2003	12.12.2019
16	nlb.ch	-	-	-	-	102	11.11.1998	13.08.2018
17	rumantsch.ch	7	586	2013	2019	188	10.05.2000	22.07.2019
18	strickcafe.ch	-	-	-	-	1’136	22.01.2009	20.10.2019
19	swiss.com	4	85’883	2014	2017	3’635	10.05.2000	22.12.2019
20	swissair.com	-	-	-	-	328	29.12.1996	10.12.2019
21	switch.ch	3	26’002	2017	2019	1’416	07.05.1997	20.11.2019
22	vigousse.ch	-	-	-	-	150	22.09.2009	14.10.2019
23	vsa-aas.ch	-	-	-	-	66	17.12.2014	02.11.2019

5.1 Analyse numérique

Sur notre échantillon de 23 sites examinés, 10 sont présents **à la BN**. La fréquence de moissonnage est généralement annuelle, à l'exception d'un site de média (#13, bisannuelle) et d'un site lié à une campagne politique (#6, une fois juste avant la votation et une fois après).

Quant à IA, nous y trouvons 22 sites sur les 23. De prime abord, il semble y avoir des nombreuses captures pour chaque cible (de quelques dizaines jusqu'à plusieurs milliers). Ce nombre indique en fait combien de fois l'URL en question a été capturée par le *crawler* d'IA. Mais certains de ces passages n'ont pas donné de résultats sous forme de pages web archivables : il peut s'agir de *redirects* ou d'autres codes réponse livrés lors de l'interrogation HTTP. Le nombre réel de versions archivées peut être par conséquent inférieur.

Pour certains sites absents de la BN il y a des explications assez évidentes (bien qu'il s'agisse de "Helvetica") :

- domaine ayant été en usage avant le début du programme Archives Web Suisse (#20),
- site privé passant « hors radar » des institutions qui proposent les URL à archiver (#5 et 18) et qui ne correspondant pas à la politique de sélection de sites web de la BN,
- entité sans ancrage dans un canton précis et qui n'entre donc pas dans le cadre des propositions effectuées par les bibliothèques cantonales (#22 et 23),
- site à but lucratif, ce qui constitue un critère d'exclusion pour la BN (#14 et 15).

IA étant à l'œuvre depuis bien plus longtemps, et moissonnant tous les contenus reliés par des liens aux *seeds* de départ, ces URL figurent bien évidemment dans ses collections.

Le seul site de notre échantillon absent d'IA (#11) est consacré à une votation cantonale. Nous supposons qu'il n'y avait que très peu, voir aucun lien qui y pointait : il s'agissait d'une prise de position d'un petit groupement ; le site était peu fourni et n'a été actif que peu de temps – à fin 2019, l'URL ne fonctionne plus.

5.2 Exploration des contenus archivés

Afin de nous faire une idée sur les contenus archivés par les deux institutions, nous avons navigué, à partir de leurs interfaces respectives, dans les sites de notre échantillon, en sélectionnant deux *snapshots* temporels différents pour chaque URL présente et en nous promenant d'un lien interne à l'autre. Il ne s'agissait pas d'une vérification systématique de l'ensemble des liens à l'aide d'un protocole prédéfini, mais nous en avons visité un grand nombre par site web.

À la BN[12], nous n'avons pas constaté des lacunes ou absences de contenus (images, sous-pages, fichiers PDF…), même à des niveaux profonds. La seule exception notable est le site de Swiss (#19), pour lequel uniquement la version en allemand semble avoir été archivée ; pour les autres versions linguistiques, seul un petit nombre de pages est accessible. Pour Heidiland.com (#7), les images manquent – nous supposons que ce contenu était encapsulé dans du code dynamique, qui ne peut être capturé correctement pas le *crawler.*

[12] Nous avons consulté les sites archivés par la BN sur un poste dédié dans les locaux de la Bibliothèque cantonale et universitaire de Lausanne (site de la Riponne). Il n'était malheureusement pas possible de faire des captures d'écran lors de cette séance de navigation en raison des droits d'accès restreints du poste en question.

Quant aux liens externes, ils ne sont en général pas actifs dans les archives web de la BN, sauf si la cible figure elle-même parmi les sites sélectionnés. Dans ce cas, e-Helvetica essaye de résoudre le lien en restant dans une zone temporelle proche de celle du *snapshot* consulté. Un exemple ce sont les liens vers le CEVA[13] depuis le site de la Ville de Carouge (#2) : depuis la capture de 2012, le lien externe renvoie vers la version 2014 du site du CEVA (*snapshot* le plus ancien), tandis que depuis la version 2017 de carouge.ch, le lien pointe également vers une capture 2017 du CEVA.

Dans IA, comme indiqué dans la section 5.3, un grand nombre de *snapshots* listés dans la vue calendrier sont en fait des *redirects* – même pour les dates indiquées en bleu dans le calendrier.

Figure 18 : Capture d'écran montrant une redirection dans la *Wayback Machine*

La destination de la redirection n'est pas tout à fait claire, mais il s'agit probablement d'une capture dans une zone temporelle proche. Par ailleurs, tous les moissonnages ne concernant pas forcément la même date, les liens internes aux sites sont aussi régulièrement résolus vers une version capturée à une date différente. Cette date est toujours indiquée en haut de la page, mais les sauts dans le temps ne sont pas annoncés autrement et peuvent donc passer inaperçus lors d'une séance de navigation.

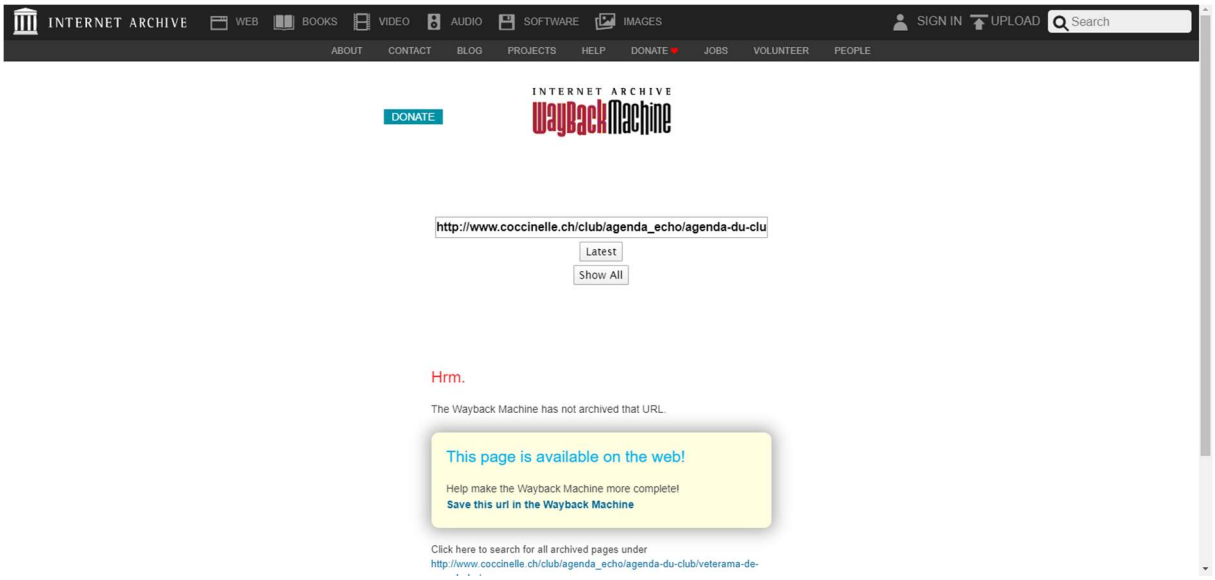
Assez régulièrement, des images sont absentes des pages archivées et remplacées par l'icône usuellement utilisée sur le web pour signaler les images manquantes (voir Figure 19).

Les sites de notre échantillon sont néanmoins bien consultables : les feuilles de style ayant été conservées, leur aspect d'époque est bien rendu. En revanche, nous avons rencontré fréquemment des liens brisés : IA indique que le contenu en question ne figure pas dans ses archives ; le cas échéant, il propose de poursuivre la navigation sur le *live web* (voir Figure 20).

[13] www.ceva.ch, site dédié au projet de liaison ferroviaire de RER dans l'agglomération genevoise (Cornavin - Eaux-Vives - Annemasse), opérationnel depuis décembre 2019 sous le nom de Léman Express. Ce site ne fait pas partie de notre échantillon, mais l'exemple nous semblait parlant.

Figure 19 : Capture d'écran d'une page archivée dont certaines images sont absentes

Figure 20 : Capture d'écran d'une URL absente de la *Wayback Machine*, mais existante dans le *live web*



Dans les deux archives du web, les champs dynamiques, à l'instar des fonctionnalités de recherche, ne sont pas actifs.

Par cette démarche, nous avons souhaité explorer une alternative pour évaluer la couverture de l'archivage du web. Mais il convient de préciser qu'il ne s'agit que d'une exploration prospective, non systématisée. Par manque de temps et de protocole nous n'avons pu collecter suffisamment de données pour formuler une analyse valide.

6. Conclusion

Le web est devenu indispensable dans notre société actuelle centrée autour de l'information et de la communication. La valeur patrimoniale d'au moins une partie de ses contenus est indiscutable. Mais il s'agit de supports volatiles et techniquement difficiles à traiter, et les volumes sont énormes.

Dans ce projet de recherche, nous nous sommes intéressées à la couverture de l'archivage du web suisse par deux acteurs, la Bibliothèque nationale suisse (BN) d'un côté et Internet Archive (IA) de l'autre. Dans un premier temps, nous avons étudié la littérature disponible sur le sujet et défini nos concepts. Notamment, le "web suisse" englobe, pour nous, les sites correspondant à la définition des "Helvetica" utilisée par la BN. Nous avons également étudié le fonctionnement organisationnel et technique des processus d'archivage du web dans les deux institutions. La différence majeure est que la BN a une approche sélective, tandis qu'IA moissonne tous les contenus rencontrés par ses crawlers, sans distinction qualitative.

Puis, pour obtenir les données brutes nécessaires à nos recherches, nous avons d'une part formulé une demande auprès de l'institution (BN), d'autre part interrogé l'API disponible à cet effet (IA). Les données brutes consistaient en des fichiers CDX, complétés par des fichiers XML en ce qui concerne la BN, avec les métadonnées de capture et de catalogage des sites moissonnés. Nous avons retravaillé ces ensembles de données à l'aide du logiciel Dataiku afin de pouvoir mener des analyses comparatives, basée sur les premières captures des domaines de premier niveau.

Nous avons souhaité compléter l'analyse quantitative par une exploration qualitative des contenus archivés. Pour ce faire, nous avons dressé une liste de 23 sites qui, selon nous, sont des "Helvetica" et avons vérifié leur présence dans les deux archives du web.

Les résultats ont montré qu'IA archive 57,44 % du ccTLD .ch, beaucoup plus largement que la BN (0.33 %). Ceci est évidemment dû au fait qu'il n'y a pas de liste définie d'URL à traiter, comme c'est le cas auprès de l'institution patrimoniale suisse. IA est également à l'œuvre depuis plus longtemps que la BN (1995 vs. 2008) et exploite davantage de *crawlers* qui opèrent en parallèle. En revanche, si un site a été sélectionné par la BN pour archivage, et selon les procédures définies par l'institution, il sera alors archivé avec un niveau qualitatif très élevé grâce à : une fréquence définie, un suivi du processus par des collaborateurs dédiés, un moissonnage de la totalité des contenus (sauf impossibilité technique), une vérification de la qualité, un référencement dans le catalogue de la bibliothèque, etc.

Un bémol au niveau de la consultation à la BN, c'est évidemment l'impossibilité d'accéder aux contenus en ligne, en dehors des institutions partenaires. Sur ce point, IA présente une meilleure expérience utilisateur, la grande majorité de ses contenus étant accessibles *via* la *Wayback machine*. Pour des raisons légales, cette pratique ne pourra probablement pas être changée de sitôt en Suisse. Mais peut-être il serait possible de publier au moins des petites images *thumbnail* sur la page de résultats de e-Helvetica Access ? Ou alors la BN pourrait faciliter l'accès aux métadonnées sur les sites archivés, selon la politique d'ouverture des données publiques qui tend à se généraliser dans les institutions publiques suisses[14], en vue d'autres projets de recherche faisant suite à celui-ci ?

[14] https://opendata.swiss/fr/

Nous pensons que les deux approches – sélectivo-qualitative d'une part, moissonnage massif fortement automatisé d'autre part – sont complémentaires. Les efforts combinés des deux acteurs examinés, ainsi que des nombreuses autres institutions qui œuvrent dans le domaine, permettront de préserver au moins une partie des contenus du web pour les générations futures.

Limitations de l'étude

Il est impossible de cerner avec exactitude ce qu'englobe le "web suisse". Nos observations et constats se réfèrent à ce que nous avons choisi de définir comme tel.

Pour les (méta-)données des deux institutions, nous étions dépendantes de ce qui est mis à disposition, soit *via* API dans le cas de IA, soit grâce à une extraction réalisée par l'équipe de la BN. Il est probable que des informations supplémentaires sont disponibles auprès des institutions, qui permettraient d'autres types d'analyses.

Dans la partie quantitative, nous nous sommes par ailleurs limitées à la comparaison au niveau du TLD, sans regarder les sous-pages des sites, ceci pour des raisons de puissance de calcul essentiellement.

Quant à l'exploration qualitative, nous avons vérifié la présence des sites de notre échantillon dans les deux archives, en nous basons sur les interfaces web plutôt que sur les API. Pour la complétude au niveau des sous-pages et autres éléments de sites, nous n'avions pas de grille d'évaluation précise, mais avons simplement noté nos observations et ressentis.

Futures recherches

Il serait intéressant de comparer les données d'autres ccTLD, issues d'IA d'une part et des différentes institutions patrimoniales nationales d'autre part, à la manière de ce que nous avons décrit dans le chapitre 4. Les fluctuations du nombre de sites moissonnés par IA d'une année à l'autre que nous avons constatées, sont-elles présentes aux mêmes moments pour d'autres pays ?

Quant à la comparaison IA-BN, on pourrait pousser plus loin la recherche qualitative et comparer de manière plus systématique encore les pages archivées pour l'échantillon de sites, en veillant à prendre des versions capturées à des moments le plus proches possibles par les deux institutions.

Bibliographie

ARCHIVE-IT, 2019a. *Archive-It - Web Archiving Services for Libraries and Archives*. [en ligne]. [Consulté le 13 janvier 2020]. Disponible à l'adresse : https://archive-it.org/

ARCHIVE-IT, 2019b. *Archive-It informational webinar - slides*.

ARCHIVES FÉDÉRALES, 2017. *Retro- und prospektiver Bewertungsentscheid betreffend Daten rund um Internet-Domain-Namen mit Endung „.ch" und ".swiss" gemäss Verordnung über Internet-Domains (VID) (ab 2015) bzw. gemäss Verordnung über die Adressierungselemente im Fernmeldebereich (AEFV) (bis 2014)* [en ligne]. 23 février 2017. [Consulté le 13 janvier 2020]. Disponible à l'adresse : https://www.bar.admin.ch/dam/bar/fr/dokumente/bewertungsentscheide/BAKOM%20Bewertungsentscheid%20Internet-Domain-Namen%20.ch%20und%20.swiss%202017.pdf

ARVIDSON, Allan, 2002. The Collection of Swedish web pages at the Royal Library - The Web Heritage of Sweden. *68th IFLA Council and General Conference, Glasgow, August 18-24, 2002* [en ligne]. La Haye : IFLA. [Consulté le 13 janvier 2020]. Disponible à l'adresse : http://archive.ifla.org/IV/ifla68/papers/111-163e.pdf

AUBRY, Sara, 2010. Introducing Web Archives as a New Library Service: the Experience of the National Library of France. *LIBER Quarterly*. 29 septembre 2010. Vol. 20, n° 2, p. 179-199. DOI 10.18352/lq.7987.

BARRETT, Katie, 2017. Pineapple Fund Gifts $1M in Bitcoin to the Internet Archive!. *Internet Archive Blogs* [en ligne]. 26 décembre 2017. [Consulté le 13 janvier 2020]. Disponible à l'adresse : https://blog.archive.org/2017/12/26/pineapple-fund-pledges-1m-in-bitcoin-to-the-internet-archive

BEAUSIRE, Jonas, 2015. *L'archivage du web: stratégies, études de cas et recommandations* [en ligne]. Genève : Haute école de gestion de Genève. Travail de bachelor. [Consulté le 13 janvier 2020]. Disponible à l'adresse : http://doc.rero.ch/record/257793

BERČIČ, Boštjan, 2005. Protection of Personal Data and Copyrighted Material on the Web: The Cases of Google and Internet Archive. *Information & Communications Technology Law*. Mars 2005. Vol. 14, n° 1, p. 17-24. DOI 10.1080/1360083042000325283.

BIBLIOTHÈQUE NATIONALE SUISSE, 2018a. FAQ sur l'archivage web. *Bibliothèque nationale suisse* [en ligne]. [Consulté le 13 janvier 2020]. Disponible à l'adresse : https://www.nb.admin.ch/snl/fr/home/informations-professionnels/e-helvetica/archives-web-suisse.html

BIBLIOTHÈQUE NATIONALE SUISSE, 2018b. Helveticat : mode d'emploi pour la recherche de sites web. *Bibliothèque nationale suisse* [en ligne]. Novembre 2018. [Consulté le 13 janvier 2020]. Disponible à l'adresse : https://www.nb.admin.ch/snl/fr/home/informations-professionnels/e-helvetica/archives-web-suisse.html

BIBLIOTHÈQUE NATIONALE SUISSE, 2018c. Tutoriel Annonce de sites web. *Bibliothèque nationale suisse* [en ligne]. 30 avril 2018. [Consulté le 13 janvier 2020]. Disponible à l'adresse : https://www.nb.admin.ch/snl/fr/home/informations-professionnels/e-helvetica/archives-web-suisse.html

BIBLIOTHÈQUE NATIONALE SUISSE, 2019. Access e-Helvetica. *Bibliothèque nationale suisse* [en ligne]. 2019. [Consulté le 13 janvier 2020]. Disponible à l'adresse : https://www.e-helvetica.nb.admin.ch

BLUMENTHAL, Karl-Rainer, 2018. Access Archive-It's Wayback index with the CDX/C API. *Archive-It Help Center* [en ligne]. 7 mai 2018. [Consulté le 13 janvier 2020]. Disponible à l'adresse : http://support.archive-it.org/hc/en-us/articles/115001790023-Access-Archive-It-s-Wayback-index-with-the-CDX-C-API

BONNEL, Sylvie et OURY, Clément, 2014. La sélection de sites web dans une bibliothèque nationale encyclopédique : une politique documentaire partagée pour le dépôt légal de l'internet à la BnF. *IFLA World Library and Information Congress 80th IFLA General Conference and Assembly, 16-22 August 2014* [en ligne]. Lyon : IFLA. [Consulté le 13 janvier 2020]. Disponible à l'adresse : http://library.ifla.org/998/1/107-bonnel-fr.pdf

BROWN, Adrian, 2006. *Archiving websites: a practical guide for information management professionals*. London : Facet Publ. ISBN 978-1-85604-553-7.

BRÜGGER, Niels, 2009. Website history and the website as an object of study. *New Media & Society*. Février 2009. Vol. 11, n° 1-2, p. 115-132. DOI 10.1177/1461444808099574.

BRUNNER, Marc-Andrea, 2014. *Internet Archive : eine urheber- und datenschutzrechtliche Analyse* [en ligne]. St. Gallen : Universität St. Gallen HSG. Masterarbeit. [Consulté le 13 janvier 2020]. Disponible à l'adresse : https://www.swissbib.ch/Record/320247902 [accès limité].

BURNS, Dasha, 2019. The Internet Archive wants to be a digital library for everything. *Sunday Closer* [en ligne]. NBC News Now. 31 mars 2019. [Consulté le 13 janvier 2020]. Disponible à l'adresse : https://www.today.com/video/the-internet-archive-wants-to-be-a-digital-library-for-everything-1468681283843

CENTRE DE COORDINATION POUR L'ARCHIVAGE À LONG TERME DE DOCUMENTS ÉLECTRONIQUES, 2013. *Qui archive le web et comment? Colloque archivage du web, 27 mai 2013* [en ligne]. [Consulté le 13 janvier 2020]. Disponible à l'adresse : https://kost-ceco.ch/cms/qui-archive-le-web-et-comment.html

CHEN, Anne, 2006. Making Web Memories with the PetaBox [en ligne]. *eWEEK*. 6 novembre 2006. p. 41-42. [Consulté le 13 janvier 2020]. Disponible à l'adresse : https://www.eweek.com/storage/making-web-memories-with-the-petabox

COSTA, Miguel, GOMES, Daniel et SILVA, Mário J., 2017. The evolution of web archiving. *International Journal on Digital Libraries*. Septembre 2017. Vol. 18, n° 3, p. 191-205. DOI 10.1007/s00799-016-0171-9.

CRESTODINA, Andy, 2017. What is the average website lifespan? 10 Factors In Website Life Expectancy. *Orbit Media Studios* [en ligne]. 25 avril 2017. [Consulté le 13 janvier 2020]. Disponible à l'adresse : https://www.orbitmedia.com/blog/website-lifespan-and-you/

CROOK, Edgar, 2009. Web archiving in a Web 2.0 world. *The Electronic Library*. 2 octobre 2009. Vol. 27, n° 5, p. 831-836. DOI 10.1108/02640470910998542.

DONIUS, Christelle et HUG BUFFO, Anna, 2019. *Revue de la littérature (bibliographie commentée). La couverture de l'archivage du web suisse : comparaison des approches de la Bibliothèque nationale suisse et de l'Internet Archive* [en ligne]. Genève : Haute école de gestion de Genève. [Consulté le 22 mai 2020]. Disponible à l'adresse : http://doc.rero.ch/record/328466/files/Donius_HugBuffo_revue_litterature.pdf

EDWARDS, Eli, 2004. Ephemeral to enduring: The Internet archive and its role in preserving digital media. *Internet Technology and Libraries* [en ligne]. Vol. 23, n° 1. [Consulté le 13 janvier 2020]. Disponible à l'adresse : https://works.bepress.com/eli_edwards/3/

GEBEIL, Sophie, 2019a. Archiver le Web, un défi historique. *The Conversation* [en ligne]. 7 juillet 2019. [Consulté le 13 janvier 2020]. Disponible à l'adresse : http://theconversation.com/archiver-le-web-un-defi-historique-117854

GEBEIL, Sophie, 2019b. Archiver les traces numériques en Méditerranée, un défi aux multiples enjeux. *The Conversation* [en ligne]. 17 juillet 2019. [Consulté le 13 janvier 2020]. Disponible à l'adresse : http://theconversation.com/archiver-les-traces-numeriques-en-mediterranee-un-defi-aux-multiples-enjeux-119041

GILL, Fiona et ELDER, Catriona, 2012. Data and archives: The Internet as site and subject. *International Journal of Social Research Methodology*. Juillet 2012. Vol. 15, n° 4, p. 271-279. DOI 10.1080/13645579.2012.687595.

GOMES, Daniel, MIRANDA, João et COSTA, Miguel, 2011. A Survey on Web Archiving Initiatives. In : GRADMANN, Stefan, BORRI, Francesca, MEGHINI, Carlo et SCHULDT, Heiko (éd.), *Research and Advanced Technology for Digital Libraries* [en ligne]. Berlin, Heidelberg : Springer Berlin Heidelberg. p. 408-420. [Consulté le 13 janvier 2020]. ISBN 978-3-642-24468-1. Disponible à l'adresse : http://link.springer.com/10.1007/978-3-642-24469-8_41

GOMES, Daniel et SILVA, Mário J., 2005. Characterizing a National Community Web. *ACM Trans. Internet Technol.* Vol. 5, n° 3, p. 508–531. DOI 10.1145/1084772.1084775.

GRAHAM, Mark, 2017. Robots.txt meant for search engines don't work well for web archives. *Internet Archive Blogs* [en ligne]. 17 avril 2017. [Consulté le 13 janvier 2020]. Disponible à l'adresse : https://blog.archive.org/2017/04/17/robots-txt-meant-for-search-engines-dont-work-well-for-web-archives/

GUY, Marieke, 2009. What's the average lifespan of a Web page? *JISC PoWR* [en ligne]. 12 août 2009. [Consulté le 13 janvier 2020]. Disponible à l'adresse : https://jiscpowr.jiscinvolve.org/wp/2009/08/12/whats-the-average-lifespan-of-a-web-page

HAKALA, Juha, 2004. Archiving the Web: European experiences. *Program*. Septembre 2004. Vol. 38, n° 3, p. 176-183. DOI 10.1108/00330330410547223.

HARDY, Quentin, 2009. Lend Ho!. *Forbes* [en ligne]. 16 novembre 2009. p. 22-24. [Consulté le 13 janvier 2020]. Disponible à l'adresse : https://www.forbes.com/forbes/2009/1116/opinions-brewster-kahle-google-ideas-opinions.html

IIPC, 2015. The CDX File Format. *WARC Specifications* [en ligne]. 2015. [Consulté le 13 janvier 2020]. Disponible à l'adresse : https://iipc.github.io/warc-specifications/specifications/cdx-format/cdx-2015/

ILLIEN, Gildas, 2008. Le Dépôt légal de l'internet en pratique : les moissonneurs du web. *Bulletin des bibliothèques de France* [en ligne]. Novembre 2008. Vol. 53, n° 6, p. 20-27. [Consulté le 13 janvier 2020]. Disponible à l'adresse : http://bbf.enssib.fr/consulter/bbf-2008-06-0020-004

ILLIEN, Gildas, 2011. Une histoire politique de l'archivage du web : le consortium international pour la préservation de l'Internet. *Bulletin des bibliothèques de France* [en ligne]. Mars 2011. Vol. 56, n° 2, p. 60-68. [Consulté le 13 janvier 2020]. Disponible à l'adresse : http://bbf.enssib.fr/consulter/bbf-2011-02-0060-012

INTERNET ARCHIVE, 2018. Heritrix 3: Glossary. *GitHub* [en ligne]. 4 juillet 2018. [Consulté le 13 janvier 2020]. Disponible à l'adresse : https://github.com/internetarchive/heritrix3/wiki/Glossary

INTERNET ARCHIVE, 2019a. About the Internet Archive. *archive.org* [en ligne]. 2019. [Consulté le 13 janvier 2020]. Disponible à l'adresse : https://archive.org/about

INTERNET ARCHIVE, 2019b. Donate to the Internet Archive! *archive.org* [en ligne]. 2019. [Consulté le 13 janvier 2020]. Disponible à l'adresse : https://archive.org/donate

INTERNET ARCHIVE, 2019c. Using The Wayback Machine. *Internet Archive Help Center* [en ligne]. 28 mai 2019. [Consulté le 13 janvier 2020]. Disponible à l'adresse : http://help.archive.org/hc/en-us/articles/360004651732-Using-The-Wayback-Machine

INTERPARES, 2020. *The InterPARES 2 Project : glossary* [en ligne]. 13 janvier 2020. [Consulté le 13 janvier 2020]. Disponible à l'adresse : http://interpares.org/ip2/display_file.cfm?doc=ip2_glossary.pdf

JavaScript Object Notation. *Wikipédia : l'encyclopédie libre* [en ligne]. Dernière modification de la page le 8 janvier 2020 à 16h43. [Consulté le 13 janvier 2020]. Disponible à l'adresse : https://fr.wikipedia.org/w/index.php?title=JavaScript_Object_Notation&oldid=166165844

KAHLE, Brewster, 1997. Preserving the Internet. *Scientific American* [en ligne]. Mars 1997. [Consulté le 13 janvier 2020]. DOI 10.1038/scientificamerican0397-82. Disponible à l'adresse : https://www.scientificamerican.com/article/preserving-the-internet [accès par abonnement]

KREYMER, Ilya, JETCRUSHERTORPEDO, BRACKETT, Rob, NEMOBIS et APONB, 2018. Wayback CDX Server API - BETA: README.md. *GitHub* [en ligne]. 1er octobre 2018. [Consulté le 13 janvier 2020]. Disponible à l'adresse : https://github.com/internetarchive/wayback/blob/master/wayback-cdx-server/README.md#basic-usage

LEETARU, Kalev, 2016. The Internet Archive Turns 20: A Behind The Scenes Look At Archiving The Web. *Forbes* [en ligne]. 18 janvier 2016. [Consulté le 13 janvier 2020]. Disponible à l'adresse : https://www.forbes.com/sites/kalevleetaru/2016/01/18/the-internet-archive-turns-20-a-behind-the-scenes-look-at-archiving-the-web/#36660cb582e0

LEPORE, Jill, 2015. The Cobweb : Can the Internet Be Archived? *The New Yorker* [en ligne]. 26 janvier 2015. [Consulté le 13 janvier 2020]. Disponible à l'adresse : https://www.newyorker.com/magazine/2015/01/26/cobweb

List of Web archiving initiatives. *Wikipédia : l'encyclopédie libre* [en ligne]. Dernière modification de la page le 11 janvier 2020 à 7h16. [Consulté le 13 janvier 2020]. Disponible à l'adresse : https://en.wikipedia.org/w/index.php?title=List_of_Web_archiving_initiatives&oldid=935217669

Liste des codes HTTP. *Wikipédia : l'encyclopédie libre* [en ligne]. Dernière modification de la page le 23 décembre 2019 à 14h35. [Consulté le 13 janvier 2020]. Disponible à l'adresse : https://fr.wikipedia.org/w/index.php?title=Liste_des_codes_HTTP&oldid=165642928

LOCHER, Hansueli, 2015. *Archives Web Suisse : Notice Archivage. Version 1.6* [en ligne]. 30 janvier 2015. Berne : Bibliothèque nationale suisse. [Consulté le 13 janvier 2020]. Disponible à l'adresse : https://www.nb.admin.ch/snl/fr/home/informations-professionnels/e-helvetica/archives-web-suisse.html

Loi fédérale du 18 décembre 1992 sur la Bibliothèque nationale suisse (Loi sur la Bibliothèque nationale, LBNS ; RS 432.21). *Les autorités fédérales de la Confédération suisse* [en ligne]. 18 décembre 1992. [Consulté le 13 janvier 2020]. Disponible à l'adresse : https://www.admin.ch/opc/fr/classified-compilation/19920349/index.html

MD5. *Wikipédia : l'encyclopédie libre* [en ligne]. Dernière modification de la page le 15 octobre 2019 à 11h14. [Consulté le 13 janvier 2020]. Disponible à l'adresse : https://fr.wikipedia.org/w/index.php?title=MD5&oldid=163556790

MINARD, Jonathan, 2013. *The Archive Documentary* [en ligne]. 10 janvier 2013. [Consulté le 13 janvier 2020]. Disponible à l'adresse :

http://archive.org/details/archive_documentary_internet_archive_sequence

MONKS-LEESON, Emily, 2011. Archives on the Internet: Representing Contexts and Provenance from Repository to Website. *The American Archivist*. 1er avril 2011. Vol. 74, n° 1, p. 38-57. DOI 10.17723/aarc.74.1.h386n333653kr83u

MUSIANI, Francesca, PALOQUE-BERGÈS, Camille, SCHAFER, Valérie et G. THIERRY, Benjamin, 2019. *Qu'est-ce qu'une archive du web ?* [en ligne]. Marseille : OpenEdition Press. [Consulté le 13 janvier 2020]. Encyclopédie numérique. ISBN 979-10-365-0470-9. Disponible à l'adresse : http://books.openedition.org/oep/8713

OFFICE FEDERAL DE LA COMMUNICATION, 2019. Statistiques [du TLD .swiss]. *nic.swiss* [en ligne]. 4 janvier 2019. [Consulté le 13 janvier 2020]. Disponible à l'adresse : https://www.nic.swiss/nic/fr/home/Hilfe/Statistiken.html

OFFICE FEDERAL DE LA COMMUNICATION, 2016. Switch. *bakom.admin.ch* [en ligne]. 27 janvier 2016. [Consulté le 13 janvier 2020]. Disponible à l'adresse : https://www.bakom.admin.ch/bakom/fr/home/glossar/switch.html

OFFICE FEDERAL DE LA COMMUNICATION, 2018. [Nouveau domaine Internet] .swiss. *bakom.admin.ch* [en ligne]. 4 avril 2018. [Consulté le 13 janvier 2020]. Disponible à l'adresse : https://www.bakom.admin.ch/bakom/fr/home/digital-und-internet/internet/internet-domain-namen/swiss-und-neue-domain-endungen-fuer-das-internet.html

Ordonnance du 14 janvier 1998 sur la Bibliothèque nationale suisse (Ordonnance sur la Bibliothèque nationale, OBNS ; RS 432.211). *Les autorités fédérales de la Confédération suisse* [en ligne]. 14 janvier 1998. [Consulté le 13 janvier 2020]. Disponible à l'adresse : https://www.admin.ch/opc/fr/classified-compilation/19980041/index.html

ORGANISATION INTERNATIONALE DE NORMALISATION, 2013. *ISO/TR 14873:2013 - Statistics and quality issues for web archiving* [en ligne]. Genève : ISO. [Consulté le 13 janvier 2020]. Disponible à l'adresse : https://www.iso.org/obp/ui/fr/#iso:std:iso:tr:14873:ed-1:v1:en [accès sur abonnement]

ORGANISATION INTERNATIONALE DE NORMALISATION, 2017. *ISO 28500:2017 - Format de fichier WARC* [en ligne]. Genève : ISO. [Consulté le 13 janvier 2020]. Disponible à l'adresse : https://www.iso.org/obp/ui/fr/#iso:std:iso:28500:ed-2:v1:en [accès sur abonnement]

OURY, Clement et POLL, Roswitha, 2013. Counting the uncountable: statistics for web archives. *Performance Measurement and Metrics*. 19 juillet 2013. Vol. 14, n° 2, p. 132-141. DOI 10.1108/PMM-05-2013-0014.

PENDSE, Liladhar R., 2016. Collecting and preserving the Ukraine conflict (2014-2015): a web archive at University of California, Berkeley. *Collection Building*. 4 juillet 2016. Vol. 35, n° 3, p. 64-72. DOI 10.1108/CB-04-2016-0006.

PENNOCK, Maureen, 2013. 13-01 : *Web-Archiving* [en ligne]. Glasgow : Digital Preservation Coalition. [Consulté le 13 janvier 2020]. DPC Technology Watch Report. Disponible à l'adresse : https://www.dpconline.org/docs/technology-watch-reports/865-dpctw13-01-pdf

RESAW, 2019. *Research infrastructure for the Study of Archived Web materials* [en ligne]. 2019. [Consulté le 13 janvier 2020]. Disponible à l'adresse : http://resaw.eu/about

SAMPATH KUMAR, B.T., VINAY KUMAR, D et PRITHVIRAJ, K.R., 2015. Wayback machine: reincarnation to vanished online citations.*Program*. 26 mars 2015. Vol. 49, n° 2, p. 205-223. DOI 10.1108/PROG-07-2013-0039.

SCHAFER, Valérie, MUSIANI, Francesca et BORELLI, Marguerite, 2016. Negotiating the Web of the Past : Web archiving, governance and STS. *French Journal for Media Research* [en ligne]. Juin 2016. n° 6/2016. [Consulté le 13 janvier 2020]. Disponible à l'adresse : http://frenchjournalformediaresearch.com/lodel-1.0/main/index.php?id=952

SHEIN, Esther, 2016. Preserving the Internet. *Communications of the ACM*. janvier 2016. Vol. 59, n° 1, p. 26-28. DOI 10.1145/2843553.

SIGNORI, Barbara, 2019a. *Archives Web Suisse : bases. Version 1.16* [en ligne]. 24 mai 2019. Berne : Bibliothèque nationale suisse. [Consulté le 13 janvier 2020]. Disponible à l'adresse : https://www.nb.admin.ch/snl/fr/home/informations-professionnels/e-helvetica/archives-web-suisse.html

SIGNORI, Barbara, 2019b. *Archives Web Suisse : Notice Collecte. Version 2.0* [en ligne]. 22 novembre 2019. Berne : Bibliothèque nationale suisse. [Consulté le 13 janvier 2020]. Disponible à l'adresse : https://www.nb.admin.ch/snl/fr/home/informations-professionnels/e-helvetica/archives-web-suisse.html

SMITH, Cathy, 2005. Building an Internet Archive System for the British Broadcasting Corporation. *Library Trends*. 2005. Vol. 54, n° 1, p. 16-32. DOI 10.1353/lib.2006.0008.

SWITCH, 2019a. Statistiques - noms de domaine. *Switch Internet Domains* [en ligne]. 2019. [Consulté le 13 janvier 2020]. Disponible à l'adresse : https://www.nic.ch/fr/statistics/domains

SWITCH, 2019b. Statistiques - part de marché. *Switch Internet Domains* [en ligne]. 2019. [Consulté le 13 janvier 2020]. Disponible à l'adresse : https://www.nic.ch/fr/statistics/market/

ULLMANN, Angela et RÖSLER, Steven, 2007. *Archivierung von Netzressourcen des Deutschen Bundestags. Version 2.0* [en ligne]. Berlin : Parlamentsarchiv des Deutschen Bundestags. [Consulté le 13 janvier 2020]. Disponible à l'adresse : https://www.bundestag.de/resource/blob/190142/e59d844a712d2d31cc66eb811650ef77/arch_netz_klein2-data.pdf

UNESCO, 2004. Charter on the Preservation of the Digital Heritage - UNESCO Bibliothèque Numérique. *Records of the General Conference, 32nd session, Paris, 29 September to 17 October 2003, v. 1: Resolutions* [en ligne]. Paris : United Nations Educational, Scientific and Cultural Organization. p. 74-77. [Consulté le 13 janvier 2020]. Disponible à l'adresse : https://unesdoc.unesco.org/ark:/48223/pf0000133171.page=80

VLCEK, Ivan, 2008. Identification and Archiving of the Czech Web Outside the National Domain. *Proceedings of the 8th International Web Archiving Workshop* [en ligne]. Aarhus, Denmark : IWAW. Septembre 2008. [Consulté le 13 janvier 2020]. Disponible à l'adresse : https://pdfs.semanticscholar.org/27a8/8fa76f6886dfd3a131c3371905bf0cbf1080.pdf

WORLDWIDEWEBSIZE.COM, 2019. *The size of the World Wide Web* [en ligne].12 janvier 2020. [Consulté le 13 janvier 2020]. Disponible à l'adresse : https://worldwidewebsize.com

Annexe 1 : Poster de recherche présenté le 12 décembre 2019

Nota bene : les chiffres et graphiques figurant en bas du poster sont erronés (voir 3.7.1 et 4.1 pour les explications). Les données correctes se trouvent en 4.1.

La couverture de l'archivage du web suisse

Comparaison des approches de la Bibliothèque nationale suisse et de Internet Archive

Contexte

Pourquoi archiver le web ?

- Documentation des pratiques culturelles de notre époque (communication, activités professionnelles, divertissement...)
- Préservation d'informations souvent uniques
- Besoins de stabilité pour : recherche, citation...

Un support particulièrement volatile qui pose des défis aux archivistes

- Structure compliquée (sous-fichiers, contenu dynamique...)
- Fluctuation et renouvellement des contenus
- Pluralité des formats

Qui sont les acteurs ? (au niveau mondial)

- Institutions patrimoniales
- Fondations
- Universités
- Médias
- Entreprises privées
- Services d'archives d'administration
- Acteurs à but lucratif

Procédé technique

- *Seed list* (URL de départ)
- *Crawler* qui suit les liens trouvés
- Profondeur du moissonnage variable
- Enregistrement p.ex. au format .WARC (web archive) avec timestamp

Notre projet de recherche

Objectifs

- Analyser et comparer deux approches différentes en matière d'archivage du web suisse
- Étudier le degré de couverture

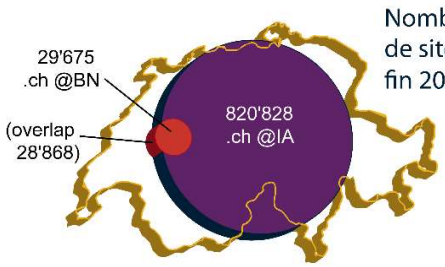
Nota bene

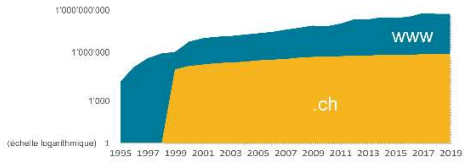
Les analyses ci-dessous s'intéressent aux sites avec extension .ch. Le « web suisse » est plus large que ce domaine de premier niveau

	Bibliothèque nationale suisse (BN)	Internet Archive (IA)
	Schweizerische Eidgenossenschaft Confédération suisse Confederazione Svizzera Confederaziun svizra	
Type	Institution patrimoniale publique	Fondation privée à but non lucratif
Siège	Berne / Suisse	San Francisco / USA
Archive le web depuis	2009	1996
Approche	Sélection de sites par des institutions partenaires	Moissonnage global, « tout ce qu'il trouve »

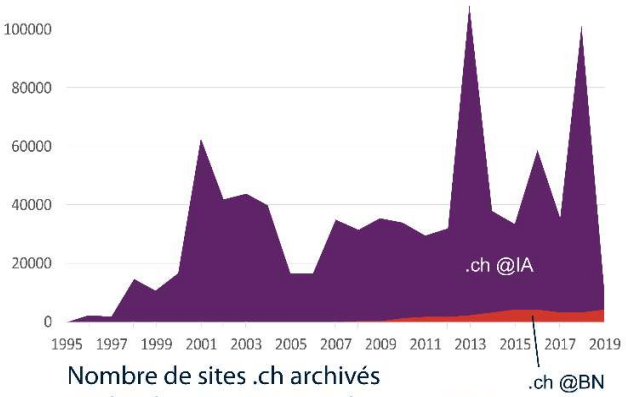
Premiers résultats

http://freevectormaps.com (carte suisse)





Part des sites .ch vs. l'ensemble du web



Nombre de sites .ch archivés par les deux institutions chaque année

h e g
Haute école de gestion Genève

Session poster du 12 décembre 2019

Projet de recherche
Master en sciences de l'information, 2018-2020

Christelle Donius – orcid.org/0000-0001-5798-3843
Anna Hug Buffo – orcid.org/0000-0001-9482-1229

Sous la direction d'Arnaud Gaudinat, professeur HES

Annexe 2 : Évolution du nombre de noms de domaines enregistrés et des nouveaux sites archivés par année

(URL retenues dans le cadre de ce projet de recherche, à savoir celles du premier niveau)

	Switch (enregistrement)		IA (archivage)		BN (archivage)	
	Nombre	**Croissance**	**Nombre**	**Croissance**	**Nombre**	**Croissance**
1995			2			
1996			2’589	129’350%		
1997			1’984	-23%		
1998			18’476	831%		
1999	175’986		14’945	-19%		
2000	162’297	-8%	23’200	55%		
2001	258’003	59%	92’664	299%		
2002	236’218	-8%	64’091	-31%		
2003	325’785	38%	65’206	2%		
2004	343’890	6%	63’040	-3%		
2005	412’525	20%	26’550	-58%		
2006	483’050	17%	26’262	-1%		
2007	568’719	18%	52’747	101%	9	
2008	668’537	18%	50’937	-3%	30	233%
2009	709’908	6%	58’454	15%	193	543%
2010	811’553	14%	54’588	-7%	421	118%
2011	842’178	4%	47’276	-13%	554	32%
2012	911’357	8%	49’487	5%	558	1%
2013	927’553	2%	177’725	259%	672	20%
2014	1’002’828	8%	54’639	-69%	667	-1%
2015	980’751	-2%	49’744	-9%	647	-3%
2016	1’058’384	8%	90’925	83%	917	42%
2017	1’070’210	1%	55’443	-39%	889	-3%
2018	1’120’441	5%	147’331	166%	1’023	15%

Annexe 3 : Répartition des TLD pour les sites “Helvetica” archivés par la BN

(URL retenues dans le cadre de ce projet de recherche, à savoir celles du premier niveau)

TLD	Nombre	%
ch	7531	92.6 %
com	341	4.2 %
org	118	1.5 %
net	48	0.6 %
info	26	0.3 %
de	15	0.2 %
swiss	8	0.1 %
eu	5	0.1 %
gr	5	0.1 %
int	5	0.1 %
li	5	0.1 %
edu	3	< 0.1 %
ag	2	< 0.1 %
cc	2	< 0.1 %
fr	2	< 0.1 %
it	2	< 0.1 %
name	2	< 0.1 %
tv	2	< 0.1 %
bg	1	< 0.1 %
bs	1	< 0.1 %
construction	1	< 0.1 %
events	1	< 0.1 %
expert	1	< 0.1 %
hockey	1	< 0.1 %
sg	1	< 0.1 %
sh	1	< 0.1 %
show	1	< 0.1 %
ws	1	< 0.1 %