

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

## **Revue de la littérature**

(Bibliographie commentée)

**réalisée par :**

**Christelle DONIUS**

**Anna HUG BUFFO**

Sous la direction de :

**Arnaud GAUDINAT, Professeur HES**

**Carouge, 6 septembre 2019**

**Master en Sciences de l'information**

**Haute École de Gestion de Genève (HEG-GE)**

# Table des matières

<b>Table des matières .....</b>	<b>1</b>
<b>Liste des abréviations .....</b>	<b>2</b>
<b>1. Introduction.....</b>	<b>3</b>
1.1 Contexte.....	3
1.2 Méthodologie.....	3
1.3 Observations préliminaires .....	4
<b>2. Revue de la littérature .....</b>	<b>5</b>
2.1 Généralités .....	5
2.2 Les bases de l'archivage du web .....	7
2.3 Initiatives nationales .....	9
2.4 Bibliothèque nationale suisse .....	12
2.5 Internet Archive .....	13
2.6 Collaborations internationales .....	14
2.7 L'incomplétude des archives du web .....	15
2.8 Aspects légaux.....	17
2.9 L'évaluation des archives du web.....	18
2.10 Définition d'un « web national » .....	18
<b>3. Bonus .....</b>	<b>19</b>
3.1 Utilisation des archives du web : exemples .....	19
3.2 Amélioration des archives du web.....	21
<b>Bibliographie .....</b>	<b>23</b>

## Liste des abréviations

BN	Bibliothèque nationale suisse
BnF	Bibliothèque nationale de France
CMS	Content Management System
ENSSIB	École nationale [française] supérieure des sciences de l'information et des bibliothèques
IA	Internet Archive
IFLA	International Federation of Library Associations and Institutions
IIPC	International Internet Preservation Consortium
MD5	Message Digest 5 (fonction de hachage cryptographique)
RGPD	Règlement général sur la protection des données (Union européenne)
TLD	Top Level Domain
ccTLD	Country Code Top Level Domain (.ch, .fr, .uk...)
gTLD	Generic Top Level Domain (.com, .org, .swiss...)
URL	Uniform Resource Locator

# 1. Introduction

## 1.1 Contexte

Notre projet de recherche vise à explorer la représentativité de l'archivage du web suisse. Plus spécifiquement, nous comparerons les approches de deux institutions, Internet Archive (IA) et la Bibliothèque nationale suisse (BN), et analyserons leurs collections selon leur complétude, leur fréquence et leur profondeur.

Afin de mener à bien cette recherche, nous avons étudié la littérature en lien avec différents sujets : définitions et indicateurs utilisés dans le domaine du web ; aspects technologiques, politiques et légaux de l'archivage de ce support spécifique ; acteurs impliqués...

Nous avons retenu une soixantaine publications, de l'article scientifique à la vidéo publiée sur Youtube en passant par la monographie, couvrant presque deux décennies d'activité dans le domaine. Nous espérons que les résumés et analyses critiques de ce document seront utiles à d'autres chercheur-e-s ou personnes intéressées par la thématique.

## 1.2 Méthodologie

Nous avons procédé à une recherche documentaire dans les catalogues de réseaux de bibliothèques RERO Explore et Swissbib, ainsi qu'auprès des ressources numériques suivantes, accessibles via l'Infothèque de la HEG :

- Emerald Insight Management Xtra
- LISTA (Library, Information Science & Technology Abstracts)
- Business Source Premier
- ERIC (Education Resource Information Center)
- Philosopher's Index

Les mots-clés que nous avons utilisés sont les suivants :

- "archive.org"
- "internet archive"
- "wayback machine"
- "web archiving"
- "webarchivierung"

Nous avons également exploité les références des articles les plus pertinents, afin d'identifier des sources supplémentaires. Notre professeur encadrant nous a par ailleurs fourni sa bibliographie Zotero sur le sujet, qu'il avait commencé à alimenter quelques années auparavant.

En complément des articles parus dans la presse spécialisée, nous avons consulté les sites web des deux institutions qui nous intéressaient particulièrement dans le cadre de ce projet de recherche, à savoir celui de la BN – dans sa partie dédiée aux e-Helvetica –, et celui d'IA, plus spécifiquement la section généraliste et le blog.

Tous les résultats pertinents ont été listés dans une grille de lecture partagée afin de pouvoir les sélectionner, les catégoriser et d'en donner une description sommaire. Les références complètes ont été gérées par Zotero.

### **1.3 Observations préliminaires**

Très vite, nous avons constaté qu'une majorité des articles datait de la période de 1996 à 2006 environ. Cela correspond au « premier essor » de l'archivage du web, avec un nombre non négligeable d'initiatives nationales ou internationales amorcées plus ou moins en même temps. Ensuite, les archives du web ayant atteint une certaine taille de corpus, des chercheurs ont commencé à les utiliser comme « matière première » et ont ensuite publié des articles traitant de l'exploitation de ces ressources documentaires. Plus récemment, une phase de maturité grandissante de la démarche s'est ouverte, avec des procédures bien établies dans les initiatives nationales et l'émergence de quelques acteurs à but lucratif ; on trouve donc des ouvrages de fond, retraçant les deux premières décennies de l'archivage du web, et des articles présentant plus de recul. Néanmoins, l'évolution technologique ne s'arrête pas, et les différents acteurs impliqués dans la constitution et l'utilisation des archives du web trouveront toujours de nouveaux thèmes propices à la rédaction de publications.

Les articles scientifiques sont généralement rédigés en anglais. Quelques publications sont en français ou en allemand. Pour les documents de travail publiés par la BN nous avons retenu la version française.

## 2. Revue de la littérature

*Ci-après, nous listons des publications, tous types de supports confondus, qui nous semblent intéressantes dans le contexte de l'archivage du web en général et du sujet de notre travail de recherche en particulier. Elles sont classées par thème, puis par ordre chronologique ; pour les rubriques qui s'y prêtent, le ou les sites web de l'institution concernée figurent en tête. En annexe se trouve une bibliographie globale par ordre alphabétique d'auteurs.*

### 2.1 Généralités

*Comme chaque phénomène de société, Internet et le web sont devenus un objet de recherche. Il y a notamment les humanités numériques qui s'y intéressent et qui produisent des réflexions plus ou moins conceptuelles à leur sujet.*

---

BRÜGGER, Niels, 2009. Website history and the website as an object of study. *New Media & Society* [en ligne]. 1er février 2009. Vol. 11, n° 1-2, p. 115-132. [Consulté le 30 août 2019]. Disponible à l'adresse : <https://doi.org/10.1177/1461444808099574> [accès par abonnement]

L'auteur est historien des médias et s'intéresse particulièrement à Internet et à ses archives. Dans cet article il essaye de définir le concept de « site web » afin que celui-ci puisse faire l'objet délimité d'une recherche. Selon lui, pour qu'un ensemble d'éléments (textes, image, etc.) sur le web constitue un site, il faut pouvoir répondre par l'affirmative aux trois questions suivantes : traitent-ils du même sujet (cohésion sémantique) ? Ont-ils un air semblable (cohésion formelle) ? Peut-on naviguer d'une fenêtre à l'autre (cohérence physique) ? Par ailleurs, il caractérise cinq strates du web à analyser : le www dans son ensemble, une sphère web (plusieurs sites liés à un même thème ou concept – ce que par exemple nous voudrions qualifier de « web suisse »), un site web, une page web, un élément d'une page web. Le danois élabore ensuite au sujet des sites web archivés que ceux-ci sont d'une part une reconstruction subjective, créée activement, et d'autre part presque toujours déficients, pour des questions techniques ou liées aux mises à jour. En effet, ces sites-là n'existaient pas, avant l'acte d'archivage, dans la forme sous laquelle ils sont désormais consultés ; ils ne sont pas le reflet exact d'un moment donné, mais plutôt un continuum temporel aux contours plus ou moins flous.

---

GILL, Fiona et ELDER, Catriona, 2012. Data and archives: The Internet as site and subject. *International Journal of Social Research Methodology* [en ligne]. 1er juillet 2012. Vol. 15, n° 4, p. 271-279. [Consulté le 30 août 2019]. Disponible à l'adresse : <https://doi.org/10.1080/13645579.2012.687595> [accès par abonnement]

Cet article pourrait presque être qualifié de philosophique. Les auteures australiennes expliquent que la notion d'archive a radicalement changé ces dernières décennies : les documents éphémères et de la vie quotidienne en font désormais partie et ouvrent des nouvelles pistes d'intérêt aux sciences sociales. Internet est utilisé à la fois comme média de communication, comme lieu d'archivage et comme objet de recherche. On peut ainsi plus facilement explorer les usages et habitudes des « gens ordinaires ». Mais il ne faut pas négliger les enjeux de protection de la vie privée : les chercheurs ont une responsabilité éthique lorsqu'ils exploitent les données publiées en ligne et doivent prendre en considération le contexte.

---

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]. N° 6/2016. [Consulté le 30 août 2019]. Disponible à l'adresse : <http://frenchjournalformediaresearch.com/lodel-1.0/main/index.php?id=952>

Les trois chercheuses démontrent comment les approches développées par les *Science and technology studies* peuvent s'appliquer à l'étude des archives du web. Elles sont à la fois la source et l'objet de ces études. Mais pour bien en saisir toute la portée, il faut comprendre l'infrastructure du web, notamment les interfaces humain-machine ; c'est ce que les auteures appellent les « négociations ». Elles font le tour des différentes questions qui préoccupent les acteurs de l'archivage du web, de la temporalité à la gouvernance. À côté d'exemples internationaux et français, elles évoquent également l'approche de la Bibliothèque nationale suisse – cas rare dans les articles scientifiques que nous avons pu consulter.

---

SUMMERS, Edward, 2019. Appraisal Practices in Web Archives. *SocArXiv* [en ligne]. 15 mars 2019. [Consulté le 30 août 2019]. Disponible à l'adresse : <https://doi.org/10.31235/osf.io/75mjp>

Ce chercheur de l'Université de Maryland collabore au projet *Documenting the Now* qui vise à rendre possible l'archivage du web et des médias sociaux par des communautés citoyennes, de manière éthique. Dans son article, il donne une vue d'ensemble des théories d'évaluation archivistique et les décline pour le web spécifiquement. Il évoque le concept de « gouvernamentalité » forgé par Michel Foucault pour indiquer le pouvoir inhérent aux décisions sur ce qui sera conservé aux archives. Par conséquent, la société démocratique elle-même devrait effectuer l'évaluation (ou du moins y participer).

---

GEBEIL, Sophie, 2019a. Archiver le Web, un défi historique. *The Conversation* [en ligne]. 7 juillet 2019. [Consulté le 30 août 2019]. 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 30 août 2019]. Disponible à l'adresse : <http://theconversation.com/archiver-les-traces-numeriques-en-mediterranee-un-defi-aux-multiples-enjeux-119041>

Dans ces deux articles, l'auteure, historienne, donne un exemple concret de l'utilisation des « sources web » par les humanités numériques. Dans le cadre de son récent doctorat, elle s'intéressait aux « mémoires de l'immigration maghrébine » sur le web. Elle a pu trouver des informations, relayant le point de vue français, dans les archives web conservées à la BnF ; mais il était nettement plus difficile d'accéder aux contenus issus du web algérien, tunisien ou marocain. Par ailleurs, durant les révolutions arabes, le web social jouait un rôle important. Mais qui archive ces sources primordiales pour les historiens du futur ?

## 2.2 Les bases de l'archivage du web

*Les bonnes pratiques pour maîtriser techniquement et intellectuellement ce support spécifique.*

---

MASANÈS, Julien (éd.), 2006. *Web archiving*. Berlin : Springer. ISBN 978-3-540-23338-1

Ouvrage de référence, cité dans de nombreux articles. Édité par l'ancien responsable de l'archivage du web à la BnF et coordinateur de l'IIPC, il rassemble des contributions de bibliothécaires et d'informaticiens. Des sujets tels que la sélection, la préservation à long terme et l'accès aux ressources archivées sont couverts, de même que l'historique des initiatives d'archivage du web, avec des exemples d'études faites sur la base de ces sources. Les choses ont évidemment évolué depuis la publication de cet ouvrage, mais beaucoup de points restent valides et il faut en tenir compte pour la gestion des archives du web.

---

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

Ce guide propose des méthodes et outils pour une mise en place pratique. Sa période de rédaction est un peu ancienne, donc le contenu n'est pas forcément à jour sur le plan technologique. Mais la couverture des différentes thématiques dont il faut tenir compte est très complète : sélection des sites à archiver, questions de droit d'auteur... Le point de vue traité est surtout anglo-saxon (Royaume-Uni, Australie), avec des mentions d'initiatives d'autres pays.

---

CHAIMBAULT, Thomas, 2008. *L'archivage du web : dossier documentaire* [en ligne]. Villeurbanne : Enssib. [Consulté le 30 août 2019]. Disponible à l'adresse : <https://www.enssib.fr/bibliotheque-numerique/notices/1730-l-archivage-du-web>

Dans ce dossier documentaire élaboré pour l'ENSSIB, on trouve des informations sur le pourquoi et le comment de l'archivage du web. L'auteur donne une typologie des différentes approches possibles et liste quelques exemples d'initiatives, tant françaises qu'internationales. La lecture de ce dossier n'apporte pas forcément davantage que Brown (2006) et Masanès (2006), mais il a l'avantage (pour les non-anglophones) d'être rédigé en français.

---

FARRELL, Susan (éd.), 2010. *A guide to web preservation: practical advice for web and records managers based on best practices from the JISC-funded PoWR project*. Oxted : Susan Farrell Consulting. ISBN 978-0-9516856-7-9

Le JISC (anciennement "Joint Information Systems Committee") est une société à but non lucratif britannique qui promeut les technologies numériques dans la formation post-obligatoire et la recherche. PoWR est l'abréviation de "preservation of web resources". Cet ouvrage s'adresse aux institutions de formation et hautes écoles, avec une approche très pratique, plus synthétique que Brown (2006). Plusieurs fiches d'études de cas (« devoirs ») listent les questions à évaluer et des pistes de solutions potentielles. L'accent est mis sur l'archivage de la ressource en tant que telle, pas sur celui du site web : si le document existe ailleurs, pas besoin de conserver la page.



---

PENNOCK, Maureen, 2013. 13-01 : *Web-Archiving* [en ligne]. Glasgow : Digital Preservation Coalition. [Consulté le 30 août 2019]. DPC Technology Watch Report. Disponible à l'adresse : <https://www.dpconline.org/docs/technology-watch-reports/865-dpctw13-01-pdf/file>

La *Digital Preservation Coalition (DPC)*, fondée en 2002 par différentes institutions patrimoniales du Royaume-Uni et de l'Irlande, œuvre pour la conservation à long terme de contenus numériques, notamment en publiant des rapports sur les bonnes pratiques d'archivage pour différents types de documents électroniques. Celui-ci sur les archives du web a été approuvé par l'IIPC. Il livre une bonne vue d'ensemble des points à considérer en la matière et des outils et logiciels disponibles. Les études de cas présentées révèlent différentes approches possibles. Une version synthétisée du rapport figure dans le *Handbook* de la DPC (2015).

---

DIGITAL PRESERVATION COALITION, 2015. *Digital Preservation Handbook : Web-archiving* [en ligne]. Glasgow : Digital Preservation Coalition. [Consulté le 30 août 2019]. Disponible à l'adresse : <https://dpconline.org/handbook/content-specific-preservation/web-archiving>

Ce chapitre du *Handbook* électronique consacré à la préservation de données numériques est basé sur le rapport de Pennock (2013). Il parle de manière succincte des différents aspects de l'archivage du web et présente quelques solutions techniques, dont notamment des offres commerciales ; ces dernières sont, selon les auteurs, un signe de la maturité du domaine. Après trois études de cas (UK Web Archive, Internet Memory Foundation et Coca-Cola Web Archive), une liste de références propose d'autres ressources (dont des vidéos sur Youtube) idéales pour les personnes relativement novices au sujet.

---

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 30 août 2019]. Disponible à l'adresse : <http://doc.rero.ch/record/257793?ln=fr>

Ce travail de bachelor en information documentaire présente un panorama des différentes stratégies que l'on peut trouver dans le domaine de l'archivage du web. Il compare plus spécifiquement la façon de procéder des deux bibliothèques nationales de la Suisse et de la France, notamment au regard du cadre législatif spécifique de chacun des pays. L'auteur a mené des entretiens avec plusieurs personnes impliquées dans ces programmes afin de compléter les informations identifiées dans la littérature. Il explore également les attentes des chercheurs quant aux archives du web. Il conclut par des réflexions conceptuelles et des défis futurs, tirés de différentes études.

---

UK WEB ARCHIVE, 2015. *What is a web archive?* [enregistrement vidéo]. Youtube [en ligne]. 2 avril 2015. [Consulté le 30 août 2019]. Disponible à l'adresse : <https://www.youtube.com/watch?v=ubDHY-ynWio>

Cette courte vidéo, faite d'illustrations, explique en anglais de manière vulgarisée la raison d'être des archives du web, et donne un aperçu des collections constituées par les institutions au Royaume-Uni.

---

POST, Colin, 2017. Building a Living, Breathing Archive: A Review of Appraisal Theories and Approaches for Web Archives. *Preservation, Digital Technology & Culture* [en ligne]. 6 janvier 2017. Vol. 46. [Consulté le 30 août 2019]. Disponible à l'adresse :

[https://www.researchgate.net/publication/319567634\\_Building\\_a\\_Living\\_Breathing\\_ArchiveA\\_Review\\_of\\_Appraisal\\_Theories\\_and\\_Approaches\\_for\\_Web\\_Archives](https://www.researchgate.net/publication/319567634_Building_a_Living_Breathing_ArchiveA_Review_of_Appraisal_Theories_and_Approaches_for_Web_Archives) [accès par abonnement]

Ce méta-article passe en revue différentes méthodes d'évaluation à appliquer aux sites web à archiver ou non. Par extension, il énumère toutes les questions à se poser lors de la mise en place d'une telle collection, de la sélection aux statistiques d'utilisation. Selon l'auteur, il faut créer des théories d'évaluation archivistiques générales, englobant tout type de support.

---

MUSIANI, Francesca, PALOQUE-BERGÈS, Camille, SCHAFER, Valérie et THIERRY, Benjamin G., 2019. *Qu'est-ce qu'une archive du web ?* [en ligne]. Marseille : OpenEdition Press. [Consulté le 30 août 2019]. Encyclopédie numérique. ISBN 979-10-365-0470-9. Disponible à l'adresse : <http://books.openedition.org/oepp/8713>

Cet ouvrage paru en Open Access se veut un aperçu très global de la thématique : les auteurs parlent à la fois de l'historique, des collaborations internationales, des approches techniques possibles, des spécificités des réseaux socio-numériques, des questions juridiques... Si l'on cherche à s'informer de manière approfondie mais générale sur le sujet, ce livre très récent est idéal. Il y a d'ailleurs une ouverture vers des théories sur les médias ou la communication. La bibliographie non plus ne se limite pas aux aspects techniques de l'archivage, mais cite des ouvrages issus des humanités numériques. En revanche, les thèmes traités sont quelque peu mélangés, les titres des chapitres ne sont pas évocateurs et il manque un index. Il est difficile de s'informer brièvement sur un point spécifique, la lecture intégrale de l'ouvrage est presque obligatoire.

## 2.3 Initiatives nationales

*Presque chaque bibliothèque nationale a publié au moins un article pour montrer qu'elle aussi, elle archive « son » web... Les articles listés ici ne représentent donc qu'une petite sélection de pays, soit parce qu'ils sont des interlocuteurs privilégiés de la Suisse sur le sujet, soit en raison de leur rôle de pionnier en la matière. Nous avons également indiqué des références concernant des vues d'ensemble des archives du web.*

---

ARVIDSON, Allan, 2002. The Collection of Swedish web pages at the Royal Library — The Web Heritage of Sweden. In : INTERNATIONAL FEDERATION OF LIBRARY ASSOCIATIONS AND INSTITUTIONS. *68th IFLA Council and General Conference, Glasgow, August 18-24, 2002* [en ligne]. La Haye : IFLA, 2002. [Consulté le 30 août 2019]. Disponible à l'adresse : <http://archive.ifla.org/IV/ifla68/papers/111-163e.pdf>

Ce compte-rendu d'une intervention au Congrès IFLA de 2002 est très sommaire, mais néanmoins intéressant pour deux aspects. D'une part, la Suède fait partie des pionniers de l'archivage du web : elle a commencé la collecte en 1997. D'autre part, l'auteur semble faire preuve de dons de prophète lorsqu'il évoque les futurs défis pour les professionnels en mentionnant les commandes vocales et la concentration des activités du web aux mains de quelques acteurs géants...

---

HAKALA, Juha, 2004. Archiving the Web: European experiences. *Program* [en ligne]. 1er septembre 2004. Vol. 38, n° 3, p. 176-183. [Consulté le 30 août 2019]. Disponible à l'adresse : <https://www.emeraldinsight.com/doi/full/10.1108/00330330410547223> [accès par abonnement]

Le directeur des technologies d'information de la Bibliothèque de l'Université de Helsinki raconte les premières années d'archivage du web en Europe, dans le cadre du projet NEDLIB (Networked European Deposit Library) financé par l'Union européenne, et englobant des partenaires en Allemagne, Finlande, France, Italie, Norvège, Pays-Bas, Portugal et Suisse. Il mentionne les défis techniques (création de processus automatisés) et politiques (nécessité d'encadrer l'activité des bibliothèques nationales par des lois sur le dépôt légal englobant les sites web, sans limitation à un TLD défini). Le fonctionnement d'un *harvester* (moissonneur) est expliqué ; en effet, NEDLIB en a développé un et a rencontré quelques difficultés techniques dans le processus. Ils utilisent le *checksum* MD5 pour éviter les doublons dans les pages archivées et garantir leur authenticité. Hakala explique ensuite les choix adoptés pour l'indexation des archives dans le cadre du projet NWA (Nordic Web Archive) des bibliothèques nationales de l'Europe du nord.

---

ULLMANN, Angela et RÖSLER, Steven, 2007. *Archivierung von Netzressourcen des Deutschen Bundestags. Version 2.0* [en ligne]. Berlin : Parlamentsarchiv des Deutschen Bundestags. [Consulté le 30 août 2019]. Disponible à l'adresse : [https://www.bundestag.de/resource/blob/190142/e59d844a712d2d31cc66eb811650ef77/arch\\_netz\\_klein2-data.pdf](https://www.bundestag.de/resource/blob/190142/e59d844a712d2d31cc66eb811650ef77/arch_netz_klein2-data.pdf)

Les archives du parlement allemand (*Bundestag*) préservent depuis 2005 les contenus des différents sites web y relatifs. Contrairement à la grande majorité des autres articles listés dans cette revue de la littérature, ce document n'est donc pas rédigé du point de vue bibliothécaire, mais fait état des spécificités d'un service d'archives. Les auteurs détaillent les réflexions à l'origine de la démarche, le concept organisationnel et la solution technique mise en place, dans le but explicite de partager leurs connaissances avec d'autres institutions. Tous les contenus de <http://webarchiv.bundestag.de> (sauf ceux en provenance de l'Intranet) sont consultables en ligne – ils l'étaient en fait déjà avant leur archivage, et l'institution détient les droits d'auteur.

---

CROOK, Edgar, 2009. Web archiving in a Web 2.0 world. *The Electronic Library* [en ligne]. 2 octobre 2009. Vol. 27, n° 5, p. 831-836. [Consulté le 30 août 2019]. Disponible à l'adresse : <https://www.emeraldinsight.com/doi/full/10.1108/02640470910998542> [accès par abonnement]

Cet article décrit la situation de la bibliothèque nationale en Australie en 2007, surtout au regard des applications du web 2.0, alors qu'elle moissonne directement une sélection de sites (avec demande de permission au propriétaire du site). Depuis 2005, un contrat la lie à IA pour un "*complete direct crawl*" du ccTLD .au. De plus, le service Archive-It (proposé également par IA, voir section 2.5) est utilisé pour ajouter des sites à la collection qui proviennent d'autres TLD, mais qui sont néanmoins jugés importants pour l'histoire australienne.

---

AUBRY, Sara, 2010. Introducing Web Archives as a New Library Service: The Experience of the National Library of France. *LIBER Quarterly* [en ligne]. 29 septembre 2010. Vol. 20, n° 2, p. 179-199. [Consulté le 30 août 2019]. Disponible à l'adresse : <http://www.liberquarterly.eu/articles/10.18352/lq.7987/>

La cheffe de projet « archivage du web » de la BnF présente la démarche de son institution. Elle détaille quatre aspects principaux : la constitution de la collection (sélection, moissonnage, cadre légal) ; la consultation par les utilisateurs finaux (recherche par URL ou par mot-clé, collections thématiques) ; les statistiques d'utilisation (liste des termes recherchés, exploitation des données des usagers collectées lors de l'inscription, enquête pour cerner les besoins en termes de fonctionnalités) ; l'implication de la communauté de bibliothécaires. Sur ce dernier point, il s'agit de faire comprendre que les collections web, même si elles sont constituées de manière automatisée, font partie du patrimoine à transmettre. Par conséquent, les responsables d'un domaine sélectionnent à la fois les imprimés et les sites web à conserver.

---

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 : Springer, p. 408-420. [Consulté le 30 août 2019]. ISBN 978-3-642-24468-1. Disponible à l'adresse : [http://link.springer.com/10.1007/978-3-642-24469-8\\_41](http://link.springer.com/10.1007/978-3-642-24469-8_41) [accès par abonnement]

Ces chercheurs portugais ont mené une enquête par e-mail au niveau international pour appréhender le nombre et les caractéristiques des différentes archives du web existantes. Celles-ci se trouvent majoritairement dans les pays développés. Les métriques collectées concernent notamment le volume des données moissonnées et leur format d'archivage, ainsi que les ressources humaines consacrées à ces activités.

---

COSTA, Miguel, GOMES, Daniel et SILVA, Mário J., 2017. The evolution of web archiving. *International Journal on Digital Libraries* [en ligne]. 1er septembre 2017. Vol. 18, n° 3, p. 191-205. [Consulté le 30 août 2019]. Disponible à l'adresse : <https://doi.org/10.1007/s00799-016-0171-9> [accès par abonnement]

La suite de l'enquête de Gomes, Miranda et Costa (2011). L'équipe, cette fois-ci, n'a pas contacté directement les acteurs de l'archivage du web, mais s'est basée sur les informations disponibles en ligne pour évaluer les changements intervenus entre 2010 et 2014. Elle a également tenu compte de données d'autres enquêtes, menées en Europe et aux États-Unis respectivement. Il s'avère que le nombre d'initiatives ainsi que les volumes collectés ont augmenté, mais que les équipes qui s'en occupent restent petites. Les pays du Sud sont toujours peu représentés. Quant aux formats et outils, on observe une tendance à la normalisation. Le grand défi reste l'accès aux archives, notamment via une recherche plein texte.

---

CHOULEUR, Marie, 2018. Ils archivent le Web : l'expérience de la Bibliothèque nationale de France [enregistrement vidéo]. In : HAUTE ÉCOLE DE GESTION GENÈVE. *100ID - 100 ans de formation en information documentaire* [en ligne]. Genève : Haute école de gestion de Genève. 19 juillet 2018. [Consulté le 30 août 2019]. Disponible à l'adresse : <https://www.youtube.com/watch?v=L8t6zcBvWAK>

Enregistrement d'une conférence donnée dans le cadre des 100 ans de formation en information documentaire à Genève, à la Haute école de gestion. La cheffe du service du dépôt légal numérique de la BnF livre un aperçu de l'archivage du web français. Elle présente les acteurs

impliqués au sein de l'institution et le cadre légal, dévoile les interfaces des outils de production et de consultation, et termine par les futures possibilités de recherche dans les données.

---

List of Web archiving initiatives. *Wikipédia : l'encyclopédie libre* [en ligne]. Dernière modification de la page le 23 juillet 2019 à 15:55 [Consulté le 30 août 2019]. Disponible à l'adresse :

[https://en.wikipedia.org/w/index.php?title=List\\_of\\_Web\\_archiving\\_initiatives&oldid=886441299](https://en.wikipedia.org/w/index.php?title=List_of_Web_archiving_initiatives&oldid=886441299)

Une liste très complète des différentes initiatives, qu'elles soient nationales, académiques, transversales ou privées. Elle a été instaurée à la suite de l'enquête de Gomes, Miranda et Costa (2011) et est mise à jour de manière collaborative.

## 2.4 Bibliothèque nationale suisse

*Objet principal (avec IA) de notre projet de recherche, la BN a commencé à réfléchir à la question de l'archivage du web en 2001 et collabore avec des institutions cantonales pour la sélection des sites.*

---

BIBLIOTHÈQUE NATIONALE SUISSE, 2018a. Collecte de sites internet patrimoniaux. *Bibliothèque nationale suisse* [en ligne]. 10 décembre 2018. [Consulté le 30 août 2019]. Disponible à l'adresse :

<https://www.nb.admin.ch/snl/fr/home/sammlungen/bibliothekssammlung/websites.html>

Description sommaire de la collection, à destination du grand public.

---

BIBLIOTHÈQUE NATIONALE SUISSE, 2018b. FAQ sur l'archivage web. *Bibliothèque nationale suisse* [en ligne]. 2018. [Consulté le 30 août 2019]. Disponible à l'adresse :

<https://www.nb.admin.ch/snl/fr/home/fachinformationen/e-helvetica/webarchiv-schweiz/faq-zu-webarchivierung.html>

Informations pour les utilisateurs et webmasters sur les raisons de l'archivage des sites web et la méthode de collecte appliquée.

---

BIBLIOTHÈQUE NATIONALE SUISSE, 2019a. Archives Web Suisse. *Bibliothèque nationale suisse* [en ligne]. 6 août 2019. [Consulté le 30 août 2019]. Disponible à l'adresse :

<https://www.nb.admin.ch/snl/fr/home/fachinformationen/e-helvetica/webarchiv-schweiz.html>

Documents à l'intention des professionnels, notamment des partenaires dans les institutions cantonales qui effectuent la sélection des sites. Notices explicatives très concrètes sur la façon de sélectionner et d'annoncer un site à ajouter à la collection. Il y a également les liens vers les ateliers organisés entre 2007 et 2015 pour mettre en place les Archives Web Suisse.

---

BIBLIOTHÈQUE NATIONALE SUISSE, 2019b. e-Helvetica Access. *Bibliothèque nationale suisse* [en ligne]. [Consulté le 30 août 2019]. Disponible à l'adresse :

[https://www.e-helvetica.nb.admin.ch/search?q=&f%5Behs\\_publication\\_type%5D%5Bo%5D=webarchive](https://www.e-helvetica.nb.admin.ch/search?q=&f%5Behs_publication_type%5D%5Bo%5D=webarchive)

Attention, la page prend un long moment pour charger. Accès à l'outil de découverte e-Helvetica pour rechercher les sites web archivés par la BN et explorer les références. À noter que, pour des raisons de respect des droits d'auteur, les résultats ne peuvent être visionnés que dans les locaux des bibliothèques partenaires.

---

## 2.5 Internet Archive

*Objet principal (avec la BN) de notre projet de recherche, IA est né de l'initiative de l'américain Brewster Kahle en 1996 et s'est donné comme objectif de donner « un accès universel à tout le savoir du monde ».*

---

INTERNET ARCHIVE, 2019a. *Internet Archive: Digital Library of Free & Borrowable Books, Movies, Music & Wayback Machine* [en ligne]. [Consulté le 30 août 2019]. Disponible à l'adresse : <https://archive.org/>

Site principal d'IA, qui donne accès à l'ensemble des collections.

---

INTERNET ARCHIVE, 2019b. *Wayback Machine* [en ligne]. [Consulté le 30 août 2019]. Disponible à l'adresse : <https://web.archive.org/>

Site de la *Wayback Machine*, le moteur de recherche spécifique aux sites web archivés.

---

INTERNET ARCHIVE, 2019c. *Internet Archive Blogs* [en ligne]. [Consulté le 30 août 2019]. Disponible à l'adresse : <https://blog.archive.org/>

Blog de l'équipe d'IA, offrant des annonces techniques, des comptes-rendus de congrès et autres événements, la promotion de quelques trésors de la collection, des commentaires sur des sujets de la gouvernance d'Internet...

---

ARCHIVE-IT, [2019]. *Archive-It - Web Archiving Services for Libraries and Archives* [en ligne]. [Consulté le 30 août 2019]. Disponible à l'adresse : <https://archive-it.org/>

La branche commerciale d'IA, qui existe depuis 2006. Les institutions ayant conclu un contrat avec Archive-It peuvent créer leurs propres collections, avec la fréquence et la profondeur de moissonnage qui correspond à leurs besoins. Une recherche plein texte est disponible pour ces collections.

---

HARDY, Quentin, 2009. Lend Ho! *Forbes* [en ligne]. 16 novembre 2009. p. 22-24. [Consulté le 30 août 2019]. Disponible à l'adresse : <https://www.forbes.com/forbes/2009/1116/opinions-brewster-kahle-google-ideas-opinions.html#61f8a73c7116>

Cet article du magazine économique américain Forbes brosse le portrait de Brewster Kahle (sa biographie, ses réalisations) et explique sa philosophie (décentralisation, ouverture), en opposition avec le fonctionnement de Google (“... *a company run by lawyers, always out to see what they can get away with. We need more choice and competition than they want.*”). L'accent est mis sur les opérations de numérisation de livres pour les grandes bibliothèques et sur la technologie *Bookserver*, qui permet de consulter, d'emprunter ou d'acheter les livres scannés disponibles sur IA.



---

LEPORE, Jill, 2015. The Cobweb: Can the Internet be archived? *The New Yorker* [en ligne]. 26 janvier 2015. [Consulté le 30 août 2019]. Disponible à l'adresse : <https://www.newyorker.com/magazine/2015/01/26/cobweb>

La journaliste nous conte l'histoire de vie de Brewster Kahle, ses motivations pour fonder IA, ses contacts internationaux... Portrait assez flatteur. L'article contient également de nombreuses informations, vulgarisées, sur les enjeux et méthodes de l'archivage du web. Très agréable à lire.

---

LEETARU, Kaleb, 2016. The Internet Archive Turns 20: A Behind the Scenes Look at Archiving the Web. *Forbes* [en ligne]. 18 janvier 2016. [Consulté le 30 août 2019]. 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/#3666ocb582eo>

L'auteur, "internet entrepreneur and academic" ([https://en.wikipedia.org/wiki/Kaleb\\_Leetaru](https://en.wikipedia.org/wiki/Kaleb_Leetaru)), insiste sur la nécessité pour les chercheurs de comprendre le fonctionnement de IA pour évaluer les biais potentiels : il ne s'agit pas d'un crawl unique (comme les moteurs de recherche le font), mais de multiples types de moissonnages, à intervalles et profondeurs variables. Par ailleurs, il y a une prépondérance de sources anglophones et de l'hémisphère ouest. Quant aux robots.txt, ces fichiers associés à un site par le webmaster menaient auparavant à l'exclusion d'un site de la collection ainsi qu'à la suppression de ses versions précédentes ; aujourd'hui, IA en tient compte en ce qui concerne l'affichage public des données, mais l'archivage a quand même lieu.

---

BURNS, Dasha, 2019. The Internet Archive wants to be a digital library for everything [enregistrement vidéo]. *Sunday Closer* [en ligne]. NBC News Now. 31 mars 2019. [Consulté le 30 août 2019]. Disponible à l'adresse : <https://www.today.com/video/the-internet-archive-wants-to-be-a-digital-library-for-everything-1468681283843>

Bref reportage sur IA par NBC News, composé d'extraits d'interview de Brewster Kahle et de quelques collaborateurs, ainsi que des images documentaires prises dans la centrale à San Francisco, notamment du bâtiment, des opérations de numérisation et des serveurs. Une entrée dans les coulisses d'IA pour les néophytes.

## 2.6 Collaborations internationales

*Depuis les débuts de l'archivage du web, l'utilité de collaborer au-delà des frontières a été reconnue. L'instance la plus importante en est le Consortium international pour la préservation de l'Internet (IIPC). D'autres collaborations existent pour des sujets spécifiques.*

---

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 30 août 2019]. Disponible à l'adresse : <http://bbf.enssib.fr/consulter/bbf-2011-02-0060-012>.

Ancien conservateur en chef au service du dépôt légal numérique de la BnF, l'auteur a participé durant plusieurs années aux travaux de l'IIPC. Dans cet article truffé d'anecdotes, il raconte la création et les premières années d'existence du Consortium et son évolution d'un groupement d'ingénieurs vers une instance internationale de lobbying.

---

POTTER, Abbey, 2012. *Why Archive the Web?* [en ligne]. 18 octobre 2012. [Consulté le 30 août 2019]. Disponible à l'adresse : <https://www.youtube.com/watch?v=pU32rjTaMFE>.

Ce clip vidéo promotionnel fait le parallèle entre la disparition de documents lors de l'incendie de la bibliothèque d'Alexandrie, il y a deux millénaires, et celle de ressources du web. Plusieurs personnalités impliquées dans l'IIPC témoignent de leurs motivations et des activités du Consortium.

---

IIPC, 2018. *International Internet Preservation Consortium* [en ligne]. [Consulté le 30 août 2019]. Disponible à l'adresse : <http://netpreserve.org/>

Créé en 2003, le Consortium a pour mission de collecter, préserver et rendre accessibles les contenus d'Internet. Il rassemble des institutions patrimoniales de près de 50 pays. Les membres collaborent pour développer des outils et des standards communs. L'IIPC organise chaque année un congrès pour favoriser le partage d'expériences et de bonnes pratiques. Les présentations de ces congrès sont partiellement accessibles sur le site, ainsi que de nombreuses autres ressources, allant de logiciels utilitaires *open source* à des études de cas sur le *text mining*. La bibliographie se limite malheureusement à la période 2001-2009.

---

RESAW, 2019. Research infrastructure for the Study of Archived Web materials. *RESAW* [en ligne]. 2019. [Consulté le 30 août 2019]. Disponible à l'adresse : <http://resaw.eu/>

Ce réseau d'institutions européennes est dédié, depuis 2012, à l'exploitation académique transversale des données archivées, et financé partiellement par le programme *Horizon 2020* de l'Union européenne. Un autre exemple de collaboration internationale, mais dédié, dans ce cas, aux usages.

## 2.7 L'incomplétude des archives du web

*Les collections archivées par IA et ses pairs sont immenses, mais ne constituent qu'une petite partie des documents ayant existé sur le web. Et même si un site a fait l'objet d'un archivage, tous ses éléments ne sont pas forcément présents. Quelle peut alors être la représentativité de ces archives ?*

---

THELWALL, Mike et VAUGHAN, Liwen, 2004. A fair history of the Web? Examining country balance in the Internet Archive. *Library & Information Science Research* [en ligne]. 1er mars 2004. Vol. 26, n° 2, p. 162-176. [Consulté le 30 août 2019]. Disponible à l'adresse : <http://www.sciencedirect.com/science/article/pii/S0740818804000246> [accès par abonnement]

Les collections d'IA offrent un grand potentiel pour des recherches historiques longitudinales au sujet du web. Mais il existe des biais géographiques qui ont une influence sur la représentativité des données qu'on y trouve. Ces deux chercheurs ont examiné la couverture d'archivage de sites commerciaux issus de quatre pays (USA, Chine, Singapour, Taïwan). Il s'avère qu'il y a des grandes différences, probablement en lien avec l'âge moyen des sites d'un pays donné et le nombre moyen d'hyperliens présents. Cette inégalité n'est donc pas intentionnelle, mais due à la technologie de base du web ; néanmoins il faut en être conscient.



---

AINSWORTH, Scott G., ALSUM, Ahmed, SALAHELDEEN, Hany, WEIGLE, Michele C. et NELSON, Michael L., 2011. How Much of the Web is Archived? In : ASSOCIATION FOR COMPUTING MACHINERY. *Proceedings of the 11th Annual International ACM/IEEE Joint Conference on Digital Libraries, Ottawa, 13-17 juin 2011* [en ligne]. New York : ACM, 2011, p. 133-136. [Consulté le 30 août 2019]. Disponible à l'adresse : <http://doi.acm.org/10.1145/1998076.1998100> [accès par abonnement]

Cette étude évalue, pour des échantillons d'URL provenant de différentes sources (Open Directory Project DMOZ, Delicious, Bitly...), le pourcentage de sites présents dans les archives web (au sens large) ; la présence du lien dans les moteurs de recherche a aussi été vérifiée. Le résultat semble peu précis de premier abord : **“35-90 % of public URIs have at least one memento”**. Selon les auteurs, les facteurs importants qui influencent l'archivage d'un site sont sa publication active par des humains (p.ex. sur Delicious, service de *social bookmarking*, actif de 2003 à 2015), son optimisation pour les moteurs de recherche et bien sûr l'archivage explicite.

---

BRUNELLE, Justin F., KELLY, Mat, SALAHELDEEN, Hany, WEIGLE, Michele C. et NELSON, Michael L., 2015. Not all mementos are created equal: measuring the impact of missing resources. *International Journal on Digital Libraries* [en ligne]. 1er septembre 2015. Vol. 16, n° 3, p. 283-301. [Consulté le 30 août 2019]. Disponible à l'adresse : <https://doi.org/10.1007/s00799-015-0150-6> [accès par abonnement]

Pour des raisons techniques, certains éléments imbriqués dans des pages web ne peuvent parfois pas être moissonnés, et les sites archivés sont donc incomplets. Un groupe de spécialistes informatiques américains a cherché à mesurer l'impact de ces éléments manquants sur l'expérience utilisateurs : si une grande image manque sur une page, il est plus important que lorsqu'il s'agit d'un tout petit logo ; si une feuille de style est absente, la forme de la page change complètement. Ils ont développé un algorithme pour évaluer cet impact et l'ont appliqué à un échantillon de plus de 45'000 pages archivées dans IA. Seules 46 % de ces pages sont complètes, et plus on avance dans le temps, plus les éléments considérés comme ayant un assez grand impact sur l'expérience-utilisateurs manquent. Ainsi, les chercheurs recommandent de faire des efforts ciblés lors des moissonnages d'archives web pour remédier à ces manques.

---

HUURDEMAN, Hugo C., KAMPS, Jaap, SAMAR, Thaer, DE VRIES, Arjen P., BEN-DAVID, Anat et ROGERS, Richard A., 2015. Lost but not forgotten: finding pages on the unarchived web. *International Journal on Digital Libraries* [en ligne]. 1er septembre 2015. Vol. 16, n° 3, p. 247-265. [Consulté le 30 août 2019]. Disponible à l'adresse : <https://doi.org/10.1007/s00799-015-0153-3>

Nous savons qu'il est impossible d'archiver l'ensemble du web. Cette équipe de l'Université d'Amsterdam s'est demandé s'il était possible de trouver des informations concernant les pages perdues, en examinant les liens depuis les pages archivées et les éléments descriptifs qui y sont associés. Il s'avère qu'on peut tracer un nombre remarquable de sites ayant disparu ; les informations y relatives sont évidemment très sommaires, mais permettent néanmoins leur identification.

## 2.8 Aspects légaux

*En matière du web, le droit d'auteur et la protection des données personnelles sont une question épineuse, qui se prolonge au-delà de son archivage.*

---

BERČIČ, Boštjan, 2005. Protection of Personal Data and Copyrighted Material on the Web: The Cases of Google and Internet Archive. *Information & Communications Technology Law* [en ligne]. 1er mars 2005. Vol. 14, n° 1, p. 17-24. [Consulté le 30 août 2019]. Disponible à l'adresse : <https://doi.org/10.1080/1360083042000325283> [accès par abonnement]

Pour ce chercheur slovaque, l'indexation et la mise en cache de pages web, telles que les moteurs de recherche et les archives du web les pratiquent, violent systématiquement la législation en matière de protection des données et de droit d'auteur. Bien avant l'entrée en vigueur du RGPD, il remarque que le droit transnational européen s'applique aussi à des acteurs extérieurs. Deux questions fondamentales se posent : un index contenant des données personnelles, constitue-t-il lui-même une « donnée personnelle » ? Et la constitution d'un tel index, correspond-elle à un traitement de données personnelles au sens de la loi ? Quant à la mise en cache de pages, elle ne respecte pas le droit d'auteur et risque même d'être nuisible lorsque des données obsolètes figurent sur cet instantané. L'auteur observe qu'en Europe, on suit généralement une politique *opt-in*, tandis qu'aux États-Unis, la règle appliquée est le *opt-out* : tout est collecté et traité, à moins que quelqu'un ne s'y oppose formellement.

---

BRUNNER, Marc-Andrea, 2014. *Internet Archive: eine urheber- und datenschutzrechtliche Analyse* [en ligne]. St. Gallen : Universität St. Gallen HSG. Masterarbeit. [Consulté le 30 août 2019]. Disponible à l'adresse : <https://www.swissbib.ch/Record/320247902> [accès limité à l'Université de St-Gall]

Contrairement à ce que pourrait laisser croire son titre, ce travail de master suisse parle des archives web en général, pas seulement d'IA qui est l'un de nos sujets d'étude. L'auteur examine les aspects de droits d'auteur et de protection des données. Il arrive à la conclusion que l'archivage de sites web risque d'enfreindre la loi de plusieurs manières et propose que la législation tienne compte des initiatives d'archives web d'intérêt général afin de permettre l'exécution de leurs activités.

---

DE BAETS, Antoon, 2016. A historian's view on the right to be forgotten. *International Review of Law, Computers & Technology* [en ligne]. 2 janvier 2016. Vol. 30, n° 1-2, p. 57-66. [Consulté le 30 août 2019]. Disponible à l'adresse : <https://doi.org/10.1080/13600869.2015.1125155>

Chercheur à l'Université de Groenigen aux Pays-Bas, cet historien juxtapose deux interprétations du terme « droit à l'oubli » : le "*right to be forgotten*", donc la protection de la sphère privée d'une personne apparaissant sur le web à un moment donné, et le "*right to forget or to remember*", qui s'apparente à la liberté d'expression. Pour lui, le droit à l'oubli, tel que prévu par la législation européenne, aurait de graves conséquences sur la recherche en histoire, voir la réécriture de celle-ci. Les historiens s'intéressent de fait, de nos jours, rarement à un individu en particulier, mais exploitent plutôt des données agrégées d'un grand nombre de personnes, afin d'étudier par exemple les comportements en société ; ainsi, une limitation des informations disponibles aux « personnes publiques » empêcherait ce type de recherche. Finalement, De Baets postule que l'oubli d'un fait ou d'une personne est inacceptable en cas de violation des droits de l'homme.

## 2.9 L'évaluation des archives du web

*Tout objet de recherche doit être mesuré afin de pouvoir être comparé et étudié. Un rapport technique de l'ISO nous propose des indicateurs, mais d'autres façons de mesurer existent.*

---

OURY, Clement et POLL, Roswitha, 2013. Counting the uncountable: statistics for web archives. *Performance Measurement and Metrics* [en ligne]. 19 juillet 2013. Vol. 14, n° 2, p. 132-141. [Consulté le 30 août 2019]. Disponible à l'adresse : <https://www.emeraldinsight.com/doi/full/10.1108/PMM-05-2013-0014> [accès par abonnement]

Paru quelques mois avant la publication officielle de son objet, cet article décrit la genèse du rapport technique ISO/TR 14873, ses raisons d'être (notamment la justification des activités face aux bailleurs de fonds), son contenu. L'IIPC avait déjà travaillé sur la standardisation dès le milieu des années 2000 avec le développement du format WARC pour les fichiers d'archivage du web, devenu la norme ISO 28500. Ici, il s'agissait de trouver un vocabulaire commun, d'identifier les bonnes pratiques, d'assurer la comparabilité internationale, dans le but d'augmenter la reconnaissance de la valeur patrimoniale et académique de ce type de données.

---

ORGANISATION INTERNATIONALE DE NORMALISATION, 2013. *ISO/TR 14873:2013 : Information and documentation — Statistics and quality issues for web archiving* [en ligne]. Genève : ISO. [Consulté le 30 août 2019]. Disponible à l'adresse : <https://www.iso.org/obp/ui/fr/#iso:std:iso:tr:14873:ed-1:v1:en> [accès par abonnement]

Développé au sein de l'IIPC, ce rapport technique vise à rendre comparables les chiffres des différentes institutions par l'utilisation d'indicateurs normalisés. Les membres du groupe de travail ont choisi de ne pas créer une norme dans un premier temps, mais ce type de document ISO d'une forme un peu moins contraignante, afin de pouvoir le publier plus vite.

---

ALSUM, Ahmed, WEIGLE, Michele C., NELSON, Michael L. et VAN DE SOMPEL, Herbert, 2013. Profiling Web Archive Coverage for Top-Level Domain and Content Language. *arXiv* [en ligne]. 16 septembre 2013. [Consulté le 30 août 2019]. Disponible à l'adresse : <http://arxiv.org/abs/1309.4008>

Ces chercheurs établissent le profil de douze archives du web selon leur couverture de TLD et de langues ; leur but est d'optimiser les requêtes envoyées par le "Memento aggregator" (voir Van de Sompel 2009, section 3.2). Il s'avère qu'IA couvre le mieux, et de loin, les échantillons testés, mais que les archives nationales s'en sortent bien dans leurs domaines spécifiques (géographiques et/ou linguistiques).

## 2.10 Définition d'un « web national »

*Le sujet de notre recherche étant le « web suisse », il s'agit de définir ce concept. D'autres auteurs ont réfléchi à la question pour leurs pays respectifs.*

---

GOMES, Daniel et SILVA, Mário J., 2005. Characterizing a National Community Web. *ACM Trans. Internet Technol.* [en ligne]. Août 2005. Vol. 5, n° 3, p. 508-531. [Consulté le 30 août 2019]. Disponible à l'adresse : <http://doi.acm.org/10.1145/1084772.1084775> [accès par abonnement]

Ces chercheurs s'intéressent au web portugais. Avant de le caractériser à l'aide de diverses statistiques, ils définissent leur objet d'étude : le web portugais correspond aux pages « ayant un intérêt culturel et sociologique pour le peuple du Portugal ». Plus concrètement, ces pages doivent satisfaire l'une de ces deux conditions :

- être hébergées sur un domaine du ccTLD .pt ;
- ou être 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.

---

VLCEK, Ivan, 2008. Identification and Archiving of the Czech Web Outside the National Domain. In : EUROPEAN ARCHIVE. *Proceedings of the 8th International Web Archiving Workshop* [en ligne]. Aarhus, Denmark : IAWAW. [Consulté le 30 août 2019]. Disponible à l'adresse : <https://pdfs.semanticscholar.org/27a8/8fa76f6886dfd3a131c3371905bfocbf1080.pdf>

Dans le cadre de son travail de bachelor à l'Université de Brno, cet informaticien a développé pour la Bibliothèque nationale tchèque un module qui s'intègre au *crawler* Heritrix. Programmé en Java, WebAnalyzer examine les pages en dehors du ccTLD .cz mais atteintes 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. Le logiciel est *open source* et peut être obtenu à l'adresse [https://is.muni.cz/th/172585/fi\\_b/?lang=en](https://is.muni.cz/th/172585/fi_b/?lang=en)

### 3. Bonus

*Dans cette section, nous présentons quelques articles dont le sujet sort du cadre de notre projet de recherche, mais que nous trouvons néanmoins intéressants dans un contexte plus large. Nous espérons qu'ils seront utiles pour les auteurs d'un futur projet.*

#### 3.1 Utilisation des archives du web : exemples

*Les collections web composées, notamment par IA, sont une ressource précieuse pour les chercheurs dans le cadre d'études très diverses.*

---

HACKETT, Stephanie et PARMANTO, Bambang, 2005. A longitudinal evaluation of accessibility: higher education web sites. *Internet Research* [en ligne]. 1er juillet 2005. Vol. 15, n° 3, p. 281-294. [Consulté le 30 août 2019]. Disponible à l'adresse : <https://emeraldinsight.com/doi/full/10.1108/10662240510602690> [accès par abonnement]

Dans leur étude, ces chercheurs ont examiné l'évolution de l'accessibilité de pages web pour des personnes handicapées à travers le temps. Pour ce faire, ils se sont basés sur les captures de sites de 45 universités américaines et 22 agences gouvernementales, archivées par IA entre 1997 et 2002. Il s'avère que les sites web deviennent moins accessibles avec la complexification des technologies utilisées, malgré l'existence de standards et guidelines.

---

SADAT-MOOSAVI, Ali, ISFANDYARI-MOGHADDAM, Alireza et TAJEDDINI, Oranus, 2012. Accessibility of online resources cited in scholarly LIS journals: A study of Emerald ISI-ranked journals. *Aslib Proceedings* [en ligne]. 16 mars 2012. Vol. 64, n° 2, p. 178-192. [Consulté le 30 août 2019]. Disponible à l'adresse : <https://emeraldinsight.com/doi/full/10.1108/00012531211215196> [accès par abonnement]

Un exemple parmi de nombreux autres qui étudie la persistance des liens cités dans des articles scientifiques (en l'occurrence de quatre journaux dans le domaine de la bibliothéconomie et des sciences de l'information, de 2005 à 2008). Sur près de 3000 liens, 64 % étaient encore fonctionnels. Grâce à différentes stratégies (correction manuelle de certains URL, recherche de l'URL dans la Wayback Machine, recherche dans Google avec des mots-clés...), 95 % des articles étaient finalement accessibles. Les auteurs suggèrent aux éditeurs de veiller à diminuer cette problématique de *link rot* en utilisant des outils tels que WebCite ou le DOI system.

---

SALAHDELDEEN, Hany M. et NELSON, Michael L., 2013. Carbon Dating the Web: Estimating the Age of Web Resources. In : ASSOCIATION FOR COMPUTING MACHINERY. *Proceedings of the 22nd International Conference on World Wide Web* [en ligne]. New York : ACM, 2013. p. 1075-1082. [Consulté le 30 août 2019]. Disponible à l'adresse : <http://doi.acm.org/10.1145/2487788.2488121> [accès par abonnement]

Deux chercheurs de l'Université Old Dominion en Virginie ont développé une application web qui permet d'estimer l'âge d'un contenu en ligne en croisant différentes sources, dont les archives du web. Il suffit de coller l'URL qui nous intéresse après le chemin suivant : <http://cd.cs.odu.edu/cd/>

---

ALNOAMANY, Yasmin, ALSUM, Ahmed, WEIGLE, Michele C. et NELSON, Michael L., 2014. Who and what links to the Internet Archive. *International Journal on Digital Libraries* [en ligne]. 1er août 2014. Vol. 14, n° 3, p. 101-115. [Consulté le 30 août 2019]. Disponible à l'adresse : <https://doi.org/10.1007/s00799-014-0111-5> [accès par abonnement]

Cet article examine les usages de la Wayback Machine (par des humains ou des robots), en se basant sur les *log files* de cette dernière : langues de la page recherchée, existence sur le web actif, si c'est un *referral link* ou un accès direct... Les chercheurs constatent que les deux tiers des pages recherchées n'existent plus que sous forme archivée, et que 82 % des utilisateurs humains arrivent sur la Wayback Machine en suivant un lien.

---

BEN-DAVID, Anat, 2016. What does the Web remember of its deleted past? An archival reconstruction of the former Yugoslav top-level domain. *New Media & Society* [en ligne]. 1er août 2016. Vol. 18, n° 7, p. 1103-1119. [Consulté le 30 août 2019]. Disponible à l'adresse : <https://doi.org/10.1177/1461444816643790> [accès par abonnement]

Le premier travail de la chercheuse israélienne dédié au ccTLD .yu, disparu en 2010 suite à la dissolution de la Yougoslavie. En se basant sur des listes de liens moissonnés juste avant sa désactivation, puis en cherchant d'autres liens dans les résultats, elle interroge IA pour reconstituer cet espace web national. Une limitation devient apparente : il faut connaître un URL exact pour pouvoir consulter les captures archivées par IA ; l'historiographie du web est donc très dépendante de ce que le *live web* peut nous indiquer. Pour interpréter les résultats,

il faut par ailleurs consulter d'autres sources, notamment afin de comprendre les enjeux géopolitiques de l'époque.

---

BEN-DAVID, Anat, AMRAM, Adam et BEKKERMAN, Ron, 2018. The colors of the national Web: visual data analysis of the historical Yugoslav Web domain. *International Journal on Digital Libraries* [en ligne]. 1er mars 2018. Vol. 19, n° 1, p. 95-106. [Consulté le 30 août 2019]. Disponible à l'adresse : <https://doi.org/10.1007/s00799-016-0202-6> [accès par abonnement]

Lors de cette recherche consécutive ont été explorées, avec des méthodes d'analyse visuelle, les images (non photographiques) contenues dans les sites du ccTLD .yu. On constate notamment que l'utilisation des couleurs du drapeau national diminue avec la progression de la guerre du Kosovo et la désintégration de l'état fédéral. Cet article est également intéressant pour son chapitre "Related work", qui signale d'autres exemples de recherches menées grâce aux archives du web. Par ailleurs, le chapitre sur la méthodologie explique comment les données ont été moissonnées depuis IA.

### 3.2 Amélioration des archives du web

*De nombreuses idées et technologies ont été trouvées afin que les collections soient plus simples à constituer ou à utiliser. Reste à voir lesquelles de ces solutions s'imposeront de manière suffisamment large pour avoir un réel impact.*

---

VAN DE SOMPEL, Herbert, NELSON, Michael L., SANDERSON, Robert, BALAKIREVA, Lyudmila L., AINSWORTH, Scott et SHANKAR, Harihar, 2009. Memento: Time Travel for the Web. *arXiv* [en ligne]. 5 novembre 2009. [Consulté le 30 août 2019]. Disponible à l'adresse : <http://arxiv.org/abs/0911.1112>

Un problème fréquent lors de la consultation de pages web archivées est que les liens mènent souvent vers des pages du *live web*, et qu'on est donc confronté à des incohérences chronologiques. Cette équipe de chercheurs propose d'utiliser une fonctionnalité du protocole HTTP : la *content negotiation*, qui permet par exemple d'indiquer une préférence linguistique. Ainsi, on pourrait définir une période temporelle : pour chaque URL saisi, le navigateur interrogerait des API de différentes archives du web et afficherait seulement une version datant de cette période.

---

ALAM, Sawood, NELSON, Michael L., VAN DE SOMPEL, Herbert, BALAKIREVA, Lyudmila L., SHANKAR, Harihar et ROSENTHAL, David S. H., 2016. Web archive profiling through CDX summarization. *International Journal on Digital Libraries* [en ligne]. 1er septembre 2016. Vol. 17, n° 3, p. 223-238. [Consulté le 30 août 2019]. Disponible à l'adresse : <https://doi.org/10.1007/s00799-016-0184-4> [accès par abonnement]

La création de « profils » des différentes archives du web, indiquant leurs contenus spécifiques, permet d'interroger celles-ci de manière ciblée et ainsi d'économiser de la bande passante. Les deux cas extrêmes de profils seraient la liste complète de tous les URL (très volumineux) d'un côté, la liste des TLD contenus dans une collection de l'autre côté (très peu spécifique). Une équipe de chercheurs américains a examiné plusieurs stratégies intermédiaires – reposant sur certains éléments du nom de domaine et du chemin d'accès à la ressource web – pour la création de profils. Les vrais et faux positifs, ainsi que les vrais et faux négatifs, ont été évalués pour chaque variante.



---

BANOS, Vangelis et MANOLOPOULOS, Yannis, 2016. A quantitative approach to evaluate Website Archivability using the CLEAR+ method. *International Journal on Digital Libraries* [en ligne]. 1er juin 2016. Vol. 17, n° 2, p. 119-141. [Consulté le 30 août 2019]. Disponible à l'adresse : <https://doi.org/10.1007/s00799-015-0144-4> [accès par abonnement]

Ces deux chercheurs grecs ont développé une méthode de mesure permettant d'évaluer si un site web pourra être archivé correctement ou non. Elle combine de nombreuses métriques, allant du temps de réponse HTTP au degré de cohésion (à savoir la présence sur le même domaine que le site de tout le contenu encasté), en passant par la présence d'attributs tels que sitemaps.xml. Ce système est implémenté dans une application web disponible sur [www.archiveready.com](http://www.archiveready.com). Les webmasters peuvent ainsi déterminer quels aspects devraient être changés pour les sites dont ils ont la responsabilité afin de garantir leur pérennité dans les archives web nationales respectives. Les institutions patrimoniales, quant à elles, pourront dédier les ressources disponibles en priorité à des sites qui promettent des bons résultats d'archivage.

---

FERNANDO, Zeon Trevor, MARENZI, Ivana et NEJDL, Wolfgang, 2018. ArchiveWeb: collaboratively extending and exploring web archive collections — How would you like to work with your collections? *International Journal on Digital Libraries* [en ligne]. 1er mars 2018. Vol. 19, n° 1, p. 39-55. [Consulté le 30 août 2019]. Disponible à l'adresse : <https://doi.org/10.1007/s00799-016-0206-2> [accès par abonnement]

Ces chercheurs du centre L3S à Hanovre ont développé un système nommé ArchiveWeb qui permet le travail collaboratif sur des collections créées via le service Archive-It d'IA. Les utilisateurs ont pu exprimer leurs besoins concernant l'annotation et l'organisation en sous-ensembles, ainsi que les recommandations pour des contenus supplémentaires. Le système est fonctionnel et disponible à l'adresse <http://archiveweb.l3s.uni-hannover.de>

---

JONES, Shawn M., NELSON, Michael L. et VAN DE SOMPEL, Herbert, 2018. Avoiding spoilers: wiki time travel with Sheldon Cooper. *International Journal on Digital Libraries* [en ligne]. 1er mars 2018. Vol. 19, n° 1, p. 77-93. [Consulté le 30 août 2019]. Disponible à l'adresse : <https://doi.org/10.1007/s00799-016-0200-8> [accès par abonnement]

Une application pratique du protocole Memento, développé par Van de Sompel et al. en 2009 (voir p. 21). Les CMS de type wiki disposant d'une fonction de *versionning* intégrée, il serait relativement simple d'exploiter les informations datées d'une page pour conduire l'internaute à une version précise. Grâce à cette méthode les amateurs de séries télévisées pourront consulter un wiki dédié, même s'ils n'ont pas encore vu le dernier épisode, sans le risque de découvrir des informations de manière précoce.

## Bibliographie

Voici la bibliographie complète des documents présentés dans cette revue de la littérature. La colonne de droite indique la rubrique dans laquelle figure notre résumé.

- AINSWORTH, Scott G., ALSUM, Ahmed, SALAH ELDEEN, Hany, WEIGLE, Michele C. et NELSON, Michael L., 2011. How Much of the Web is Archived? In : ASSOCIATION FOR COMPUTING MACHINERY. *Proceedings of the 11th Annual International ACM/IEEE Joint Conference on Digital Libraries, Ottawa, 13-17 juin 2011* [en ligne]. New York : ACM, 2011, p. 133–136. [Consulté le 30 août 2019]. Disponible à l'adresse : <http://doi.acm.org/10.1145/1998076.1998100> [accès par abonnement] 2.7
- ALAM, Sawood, NELSON, Michael L., VAN DE SOMPEL, Herbert, BALAKIREVA, Lyudmila L., SHANKAR, Harihar et ROSENTHAL, David S. H., 2016. Web archive profiling through CDX summarization. *International Journal on Digital Libraries* [en ligne]. 1er septembre 2016. Vol. 17, n° 3, p. 223-238. [Consulté le 30 août 2019]. Disponible à l'adresse : <https://doi.org/10.1007/s00799-016-0184-4> [accès par abonnement] 3.2
- ALNOAMANY, Yasmin, ALSUM, Ahmed, WEIGLE, Michele C. et NELSON, Michael L., 2014. Who and what links to the Internet Archive. *International Journal on Digital Libraries* [en ligne]. 1er août 2014. Vol. 14, n° 3, p. 101-115. [Consulté le 30 août 2019]. Disponible à l'adresse : <https://doi.org/10.1007/s00799-014-0111-5> [accès par abonnement] 3.1
- ALSUM, Ahmed, WEIGLE, Michele C., NELSON, Michael L. et VAN DE SOMPEL, Herbert, 2013. Profiling Web Archive Coverage for Top-Level Domain and Content Language. *arXiv* [en ligne]. 16 septembre 2013. [Consulté le 30 août 2019]. Disponible à l'adresse : <http://arxiv.org/abs/1309.4008> 2.9
- ARCHIVE-IT, [2019]. *Archive-It - Web Archiving Services for Libraries and Archives* [en ligne]. [Consulté le 30 août 2019]. Disponible à l'adresse : <https://archive-it.org/> 2.5
- ARVIDSON, Allan, 2002. The Collection of Swedish web pages at the Royal Library — The Web Heritage of Sweden. In : INTERNATIONAL FEDERATION OF LIBRARY ASSOCIATIONS AND INSTITUTIONS. *68th IFLA Council and General Conference, Glasgow, August 18-24, 2002* [en ligne]. La Haye : IFLA, 2002. [Consulté le 30 août 2019]. Disponible à l'adresse : <http://archive.ifla.org/IV/ifla68/papers/111-163e.pdf> 2.3
- AUBRY, Sara, 2010. Introducing Web Archives as a New Library Service: the Experience of the National Library of France. *LIBER Quarterly* [en ligne]. 29 septembre 2010. Vol. 20, n° 2, p. 179-199. [Consulté le 30 août 2019]. Disponible à l'adresse : <http://www.liberquarterly.eu/articles/10.18352/lq.7987/> 2.3
- BANOS, Vangelis et MANOLOPOULOS, Yannis, 2016. A quantitative approach to evaluate Website Archivability using the CLEAR+ method. *International Journal on Digital Libraries* [en ligne]. 1er juin 2016. Vol. 17, n° 2, p. 119-141. [Consulté le 30 août 2019]. Disponible à l'adresse : <https://doi.org/10.1007/s00799-015-0144-4> [accès par abonnement] 3.2



- 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 30 août 2019]. Disponible à l'adresse : <http://doc.rero.ch/record/257793?ln=fr> 2.2
- BEN-DAVID, Anat, 2016. What does the Web remember of its deleted past? An archival reconstruction of the former Yugoslav top-level domain. *New Media & Society* [en ligne]. 1er août 2016. Vol. 18, n° 7, p. 1103-1119. [Consulté le 30 août 2019]. Disponible à l'adresse : <https://doi.org/10.1177/1461444816643790> [accès par abonnement] 3.1
- BEN-DAVID, Anat, AMRAM, Adam et BEKKERMAN, Ron, 2018. The colors of the national Web: visual data analysis of the historical Yugoslav Web domain. *International Journal on Digital Libraries* [en ligne]. 1er mars 2018. Vol. 19, n° 1, p. 95-106. [Consulté le 30 août 2019]. Disponible à l'adresse : <https://doi.org/10.1007/s00799-016-0202-6> [accès par abonnement] 3.1
- BERČIČ, Boštjan, 2005. Protection of Personal Data and Copyrighted Material on the Web: The Cases of Google and Internet Archive. *Information & Communications Technology Law* [en ligne]. 1er mars 2005. Vol. 14, n° 1, p. 17-24. [Consulté le 30 août 2019]. Disponible à l'adresse : <https://doi.org/10.1080/1360083042000325283> [accès par abonnement] 2.8
- BIBLIOTHÈQUE NATIONALE SUISSE, 2019a. Archives Web Suisse. *Bibliothèque nationale suisse* [en ligne]. 6 août 2019. [Consulté le 30 août 2019]. Disponible à l'adresse : <https://www.nb.admin.ch/snl/fr/home/fachinformationen/e-helvetica/webarchiv-schweiz.html> 2.4
- BIBLIOTHÈQUE NATIONALE SUISSE, 2019b. e-Helvetica Access. *Bibliothèque nationale suisse* [en ligne]. [Consulté le 30 août 2019]. Disponible à l'adresse : [https://www.e-helvetica.nb.admin.ch/search?q=&f%5Behs\\_publication\\_type%5D%5B0%5D=webarchive](https://www.e-helvetica.nb.admin.ch/search?q=&f%5Behs_publication_type%5D%5B0%5D=webarchive) 2.4
- BIBLIOTHÈQUE NATIONALE SUISSE, 2018a. Collecte de sites internet patrimoniaux. *Bibliothèque nationale suisse* [en ligne]. 10 décembre 2018. [Consulté le 30 août 2019]. Disponible à l'adresse : <https://www.nb.admin.ch/snl/fr/home/sammlungen/bibliothekssammlung/websites.html> 2.4
- BIBLIOTHÈQUE NATIONALE SUISSE, 2018b. FAQ sur l'archivage web. *Bibliothèque nationale suisse* [en ligne]. [Consulté le 30 août 2019]. Disponible à l'adresse : <https://www.nb.admin.ch/snl/fr/home/fachinformationen/e-helvetica/webarchiv-schweiz/faq-zu-webarchivierung.html> 2.4
- BROWN, Adrian, 2006. *Archiving websites: a practical guide for information management professionals*. London: Facet Publ. ISBN 978-1-85604-553-7 2.2
- BRÜGGER, Niels, 2009. Website history and the website as an object of study. *New Media & Society* [en ligne]. 1er février 2009. Vol. 11, n° 1-2, p. 115-132. [Consulté le 30 août 2019]. Disponible à l'adresse : <https://doi.org/10.1177/1461444808099574> [accès par abonnement] 2.1

- BRUNELLE, Justin F., KELLY, Mat, SALAHELDEEN, Hany, WEIGLE, Michele C. et NELSON, Michael L., 2015. Not all mementos are created equal: measuring the impact of missing resources. *International Journal on Digital Libraries* [en ligne]. 1er septembre 2015. Vol. 16, n° 3, p. 283-301. [Consulté le 30 août 2019]. Disponible à l'adresse : <https://doi.org/10.1007/s00799-015-0150-6> [accès par abonnement] 3.2
- BRUNNER, Marc-Andrea, 2014. *Internet Archive: eine urheber- und datenschutzrechtliche Analyse* [en ligne]. St. Gallen : Universität St. Gallen HSG. Masterarbeit. [Consulté le 30 août 2019]. Disponible à l'adresse : <https://www.swissbib.ch/Record/320247902> [accès limité à l'Université de St-Gall] 2.8
- BURNS, Dasha, 2019. The Internet Archive wants to be a digital library for everything [enregistrement vidéo]. *Sunday Closer* [en ligne]. NBC News Now. 31 mars 2019. [Consulté le 30 août 2019]. Disponible à l'adresse : <https://www.today.com/video/the-internet-archive-wants-to-be-a-digital-library-for-everything-1468681283843> 2.5
- CHAIMBAULT, Thomas, 2008. *L'archivage du web : dossier documentaire* [en ligne]. Villeurbanne : Enssib. [Consulté le 30 août 2019]. Disponible à l'adresse : <https://www.enssib.fr/bibliotheque-numerique/notices/1730-l-archivage-du-web> 2.2
- CHOULEUR, Marie, 2018. Ils archivent le Web, l'expérience de la Bibliothèque nationale de France [enregistrement vidéo]. In : HAUTE ÉCOLE DE GESTION GENÈVE. *100ID - 100 ans de formation en information documentaire* [en ligne]. Genève : Haute école de gestion de Genève. 19 juillet 2018. [Consulté le 30 août 2019]. Disponible à l'adresse : <https://www.youtube.com/watch?v=L8t6zcBvWAK> 2.3
- COSTA, Miguel, GOMES, Daniel et SILVA, Mário J., 2017. The evolution of web archiving. *International Journal on Digital Libraries* [en ligne]. 1er septembre 2017. Vol. 18, n° 3, p. 191-205. [Consulté le 30 août 2019]. Disponible à l'adresse : <https://doi.org/10.1007/s00799-016-0171-9> [accès par abonnement] 2.3
- CROOK, Edgar, 2009. Web archiving in a Web 2.0 world. *The Electronic Library* [en ligne]. 2 octobre 2009. Vol. 27, n° 5, p. 831-836. [Consulté le 30 août 2019]. Disponible à l'adresse : <https://www.emeraldinsight.com/doi/full/10.1108/02640470910998542> [accès par abonnement] 2.3
- DE BAETS, Antoon, 2016. A historian's view on the right to be forgotten. *International Review of Law, Computers & Technology* [en ligne]. 2 janvier 2016. Vol. 30, n° 1-2, p. 57-66. [Consulté le 30 août 2019]. Disponible à l'adresse : <https://doi.org/10.1080/13600869.2015.1125155> 2.8
- DIGITAL PRESERVATION COALITION, 2015. *Digital Preservation Handbook : Web-archiving* [en ligne]. Glasgow : Digital Preservation Coalition. [Consulté le 30 août 2019]. Disponible à l'adresse : <https://dpconline.org/handbook/content-specific-preservation/web-archiving> 2.2
- FARRELL, Susan (éd.), 2010. *A guide to web preservation: practical advice for web and records managers based on best practices from the JISC-funded PoWR project*. Oxted : Susan Farrell Consulting. ISBN 978-0-9516856-7-9 2.2

- FERNANDO, Zeon Trevor, MARENZI, Ivana et NEJDL, Wolfgang, 2018. ArchiveWeb: collaboratively extending and exploring web archive collections — How would you like to work with your collections? *International Journal on Digital Libraries* [en ligne]. 1<sup>er</sup> mars 2018. Vol. 19, n° 1, p. 39-55. [Consulté le 30 août 2019]. Disponible à l'adresse : <https://doi.org/10.1007/s00799-016-0206-2> [accès par abonnement] 3.2
- GEBEIL, Sophie, 2019a. Archiver le Web, un défi historique. *The Conversation* [en ligne]. 7 juillet 2019. [Consulté le 30 août 2019]. Disponible à l'adresse : <http://theconversation.com/archiver-le-web-un-defi-historique-117854> 2.1
- GEBEIL, Sophie, 2019b. Archiver les traces numériques en Méditerranée, un défi aux multiples enjeux. *The Conversation* [en ligne]. 17 juillet 2019. [Consulté le 30 août 2019]. Disponible à l'adresse : <http://theconversation.com/archiver-les-traces-numeriques-en-mediterranee-un-defi-aux-multiples-enjeux-119041> 2.1
- GILL, Fiona et ELDER, Catriona, 2012. Data and archives: The Internet as site and subject. *International Journal of Social Research Methodology* [en ligne]. 1er juillet 2012. Vol. 15, n° 4, p. 271-279. [Consulté le 30 août 2019]. Disponible à l'adresse : <https://doi.org/10.1080/13645579.2012.687595> [accès par abonnement] 2.1
- 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 : Springer, p. 408-420. [Consulté le 30 août 2019]. ISBN 978-3-642-24468-1. Disponible à l'adresse : [http://link.springer.com/10.1007/978-3-642-24469-8\\_41](http://link.springer.com/10.1007/978-3-642-24469-8_41) [accès par abonnement] 2.3
- GOMES, Daniel et SILVA, Mário J., 2005. Characterizing a National Community Web. *ACM Trans. Internet Technol.* [en ligne]. Août 2005. Vol. 5, n° 3, p. 508–531. [Consulté le 30 août 2019]. Disponible à l'adresse : <http://doi.acm.org/10.1145/1084772.1084775> [accès par abonnement] 2.10
- HACKETT, Stephanie et PARMANTO, Bambang, 2005. A longitudinal evaluation of accessibility: higher education web sites. *Internet Research* [en ligne]. 1er juillet 2005. Vol. 15, n° 3, p. 281-294. [Consulté le 30 août 2019]. Disponible à l'adresse : <https://emeraldinsight.com/doi/full/10.1108/10662240510602690> [accès par abonnement] 3.1
- HAKALA, Juha, 2004. Archiving the Web: European experiences. *Program* [en ligne]. 1er septembre 2004. Vol. 38, n° 3, p. 176-183. [Consulté le 30 août 2019]. Disponible à l'adresse : <https://www.emeraldinsight.com/doi/full/10.1108/00330330410547223> [accès par abonnement] 2.3
- HARDY, Quentin, 2009. Lend Ho! *Forbes* [en ligne]. 16 novembre 2009. p. 22-24. [Consulté le 30 août 2019]. Disponible à l'adresse : <https://www.forbes.com/forbes/2009/1116/opinions-brewster-kahle-google-ideas-opinions.html#61f8a73c7116> 2.5
- HUURDEMAN, Hugo C., KAMPS, Jaap, SAMAR, Thaer, DE VRIES, Arjen P., BEN-DAVID, Anat et ROGERS, Richard A., 2015. Lost but not forgotten: finding pages on the unarchived web. *International Journal on Digital Libraries* [en ligne]. 1er septembre 2015. Vol. 16, n° 3, p. 247-265. [Consulté le 30 août 2019]. Disponible à l'adresse : <https://doi.org/10.1007/s00799-015-0153-3> 3.2

- IIPC, 2018. *International Internet Preservation Consortium* [en ligne]. [Consulté le 30 août 2019]. Disponible à l'adresse : <http://netpreserve.org/> 2.6
- 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 30 août 2019]. Disponible à l'adresse : <http://bbf.enssib.fr/consulter/bbf-2011-02-0060-012> 2.6
- INTERNET ARCHIVE, 2019a. *Internet Archive Blogs* [en ligne]. [Consulté le 30 août 2019]. Disponible à l'adresse : <https://blog.archive.org/> 2.5
- INTERNET ARCHIVE, 2019b. *Internet Archive: Digital Library of Free & Borrowable Books, Movies, Music & Wayback Machine* [en ligne]. [Consulté le 30 août 2019]. Disponible à l'adresse : <https://archive.org/> 2.5
- INTERNET ARCHIVE, 2019c. *Wayback Machine* [en ligne]. [Consulté le 30 août 2019]. Disponible à l'adresse : <https://web.archive.org/> 2.5
- JONES, Shawn M., NELSON, Michael L. et VAN DE SOMPEL, Herbert, 2018. Avoiding spoilers: wiki time travel with Sheldon Cooper. *International Journal on Digital Libraries* [en ligne]. 1er mars 2018. Vol. 19, n° 1, p. 77-93. [Consulté le 30 août 2019]. Disponible à l'adresse : <https://doi.org/10.1007/s00799-016-0200-8> [accès par abonnement] 3.2
- 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 30 août 2019]. 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> 2.5
- LEPORE, Jill, 2015. The Cobweb: Can the Internet be archived? *The New Yorker* [en ligne]. 26 janvier 2015. [Consulté le 30 août 2019]. Disponible à l'adresse : <https://www.newyorker.com/magazine/2015/01/26/cobweb> 2.5
- List of Web archiving initiatives. *Wikipédia : l'encyclopédie libre* [en ligne]. Dernière modification de la page le 23 juillet 2019 à 15:55 [Consulté le 30 août 2019]. Disponible à l'adresse : [https://en.wikipedia.org/w/index.php?title=List\\_of\\_Web\\_archiving\\_initiatives&oldid=886441299](https://en.wikipedia.org/w/index.php?title=List_of_Web_archiving_initiatives&oldid=886441299) 2.3
- MASANÈS, Julien (éd.), 2006. *Web archiving*. Berlin : Springer. ISBN 978-3-540-23338-1 2.2
- MUSIANI, Francesca, PALOQUE-BERGÈS, Camille, SCHAFER, Valérie et THIERRY, Benjamin G., 2019. *Qu'est-ce qu'une archive du web ?* [en ligne]. Marseille : OpenEdition Press. [Consulté le 30 août 2019]. Encyclopédie numérique. ISBN 979-10-365-0470-9. Disponible à l'adresse : <http://books.openedition.org/oep/8713> 2.2
- ORGANISATION INTERNATIONALE DE NORMALISATION, 2013. *ISO/TR 14873:2013: Information and documentation — Statistics and quality issues for web archiving* [en ligne]. Genève : ISO. [Consulté le 30 août 2019]. Disponible à l'adresse : <https://www.iso.org/obp/ui/fr/#iso:std:iso:tr:14873:ed-1:v1:en> [accès par abonnement] 2.9

- OURY, Clement et POLL, Roswitha, 2013. Counting the uncountable: statistics for web archives. *Performance Measurement and Metrics* [en ligne]. 19 juillet 2013. Vol. 14, n° 2, p. 132-141. [Consulté le 30 août 2019]. Disponible à l'adresse : <https://www.emeraldinsight.com/doi/full/10.1108/PMM-05-2013-0014> [accès par abonnement] 2.9
- PENNOCK, Maureen, 2013. 13-01: *Web-Archiving* [en ligne]. Glasgow : Digital Preservation Coalition. DPC Technology Watch Report. [Consulté le 30 août 2019]. Disponible à l'adresse : <https://www.dpconline.org/docs/technology-watch-reports/865-dpctw13-01-pdf/file> 2.2
- POST, Colin, 2017. Building a Living, Breathing Archive: A Review of Appraisal Theories and Approaches for Web Archives. *Preservation, Digital Technology & Culture* [en ligne]. 6 janvier 2017. Vol. 46. [Consulté le 30 août 2019]. Disponible à l'adresse : [https://www.researchgate.net/publication/319567634\\_Building\\_a\\_Living\\_Breathing\\_ArchiveA\\_Review\\_of\\_Appraisal\\_Theories\\_and\\_Approaches\\_for\\_Web\\_Archives](https://www.researchgate.net/publication/319567634_Building_a_Living_Breathing_ArchiveA_Review_of_Appraisal_Theories_and_Approaches_for_Web_Archives) [accès par abonnement] 2.2
- POTTER, Abbey, 2012. *Why Archive the Web?* [en ligne]. 18 octobre 2012. [Consulté le 30 août 2019]. Disponible à l'adresse : <https://www.youtube.com/watch?v=pU32rjTaMFE> 2.6
- RESAW, 2019. Research infrastructure for the Study of Archived Web materials. *RESAW* [en ligne]. 2019. [Consulté le 30 août 2019]. Disponible à l'adresse : <http://resaw.eu/> 2.6
- SADAT-MOOSAVI, Ali, ISFANDYARI-MOGHADDAM, Alireza et TAJEDDINI, Oranus, 2012. Accessibility of online resources cited in scholarly LIS journals: A study of Emerald ISI-ranked journals. *Aslib Proceedings* [en ligne]. 16 mars 2012. Vol. 64, n° 2, p. 178-192. [Consulté le 30 août 2019]. Disponible à l'adresse : <https://emeraldinsight.com/doi/full/10.1108/00012531211215196> [accès par abonnement] 3.1
- SALAHDELDEEN, Hany M. et NELSON, Michael L., 2013. Carbon Dating the Web: Estimating the Age of Web Resources. In : ASSOCIATION FOR COMPUTING MACHINERY. *Proceedings of the 22nd International Conference on World Wide Web* [en ligne]. New York : ACM, 2013. p. 1075–1082. [Consulté le 30 août 2019]. Disponible à l'adresse : <http://doi.acm.org/10.1145/2487788.2488121> [accès par abonnement] 3.1
- 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]. N° 6/2016. [Consulté le 30 août 2019]. Disponible à l'adresse : <http://frenchjournalformediaresearch.com/lovel-1.0/main/index.php?id=952> 2.1
- SUMMERS, Edward, 2019. Appraisal Practices in Web Archives. *SocArXiv* [en ligne]. 15 mars 2019. [Consulté le 30 août 2019]. Disponible à l'adresse : <https://doi.org/10.31235/osf.io/75mjp> 2.1
- THELWALL, Mike et VAUGHAN, Liwen, 2004. A fair history of the Web? Examining country balance in the Internet Archive. *Library & Information Science Research* [en ligne]. 1er mars 2004. Vol. 26, n° 2, p. 162-176. [Consulté le 30 août 2019]. Disponible à l'adresse : <http://www.sciencedirect.com/science/article/pii/S0740818804000246> [accès par abonnement] 2.7



- UK WEB ARCHIVE, 2015. *What is a web archive?* [enregistrement vidéo]. *Youtube* 2.2  
[en ligne]. 2 avril 2015. [Consulté le 30 août 2019]. Disponible à l'adresse :  
<https://www.youtube.com/watch?v=ubDHY-ynWi0>
- ULLMANN, Angela et RÖSLER, Steven, 2007. *Archivierung von Netzressourcen des Deutschen Bundestags. Version 2.0* [en ligne]. Berlin : Parlamentsarchiv des Deutschen Bundestags. [Consulté le 30 août 2019]. Disponible à l'adresse :  
[https://www.bundestag.de/resource/blob/190142/e59d844a712d2d31cc66eb811650ef77/arch\\_netz\\_klein2-data.pdf](https://www.bundestag.de/resource/blob/190142/e59d844a712d2d31cc66eb811650ef77/arch_netz_klein2-data.pdf) 2.3
- VAN DE SOMPEL, Herbert, NELSON, Michael L., SANDERSON, Robert, BALAKIREVA, Lyudmila L., AINSWORTH, Scott et SHANKAR, Harihar, 2009. Memento: Time Travel for the Web. *arXiv* [en ligne]. 5 novembre 2009. [Consulté le 30 août 2019]. Disponible à l'adresse : <http://arxiv.org/abs/0911.1112> 3.2
- VLCEK, Ivan, 2008. Identification and Archiving of the Czech Web Outside the National Domain. In : EUROPEAN ARCHIVE. *Proceedings of the 8th International Web Archiving Workshop* [en ligne]. Aarhus, Denmark : IWAW. [Consulté le 30 août 2019]. Disponible à l'adresse : <https://pdfs.semanticscholar.org/27a8/8fa76f6886dfd3a131c3371905bf0cbf1080.pdf> 2.10