Vers un archivage fédéral du web

La recherche d'une stratégie pour un archivage fédéral du web

Friedel Geeraert

Le web est devenu quasi incontournable dans notre vie quotidienne. Ce nouveau média contient en conséquence de plus en plus de traces de notre histoire. La Bibliothèque royale de Belgique et les Archives de l'État analysent ensemble une stratégie commune en vue de sauvegarder le contenu du web.

Un paradoxe

Bien que le web soit très précieux d'un point de vue patrimonial, compte tenu notamment de la grande diversité de personnes et d'organisations qui publient du contenu en ligne, peu d'attention est portée sur la préservation de son contenu. Le web est en perpétuelle évolution : des informations y sont ajoutées, modifiées ou supprimées. La durée de vie relativement courte des informations sur le web a pour corollaire qu'une partie du patrimoine belge en ligne disparaît quotidiennement.

À l'étranger, de nombreux services nationaux d'archives et de bibliothèques s'occupent depuis plusieurs années de l'archivage de certaines parties de leur web national, et ces archives deviennent de plus en plus des sujets d'étude pour les chercheurs. En Belgique, il y a certes quelques initiatives de petite envergure pour archiver le web, comme par ex. au Felixarchief à Anvers ou à la bibliothèque universitaire de Gand, mais il n'existe pas encore de stratégie nationale ou fédérale. Afin de combler cela, la Bibliothèque royale (KBR) et les Archives de l'État (AGR) ont uni leurs forces pour lancer un projet de recherche sur l'archivage fédéral du web belge.

Le projet de recherche

Lancé le 1 juin 2017, le projet de recherche PROMISE (Preserving Online Multiple Information: towards a Belgian StratEgy) vise à élaborer une stratégie fédérale pour la conservation du web belge. Ce projet est très innovant dans le sens où il réunit deux établissements scientifiques fédéraux en vue de l'élaboration de collections d'archives belges du web, créant ainsi une situation « gagnant-gagnant » vu que les AGR et KBR peuvent mutualiser leurs expertises et leurs infrastructures. Pendant les deux années que durera le projet, les objectifs suivants seront visés :

1. Identification de bonnes pratiques ayant trait à l'archivage du web
2. Élaboration d'une politique belge pour l'archivage du web
3. Lancement d'un projet pilote pour l'archivage du web belge
4. Recommandations pour une implémentation d'un service d'archivage durable du web

Comment archive-t-on le web ?

Parmi les différentes façons pour archiver le web, les méthodes les plus utilisées sont le « *crawling* » (indexation) ou le « *harvesting* » (extraction) de sites internet via des « *crawler robots* ». Sur base d'une liste d'adresses de sites web (URL), le robot effectue des demandes à un serveur web, puis enregistre le contenu communiqué par le serveur, ensuite dresse une liste de tous les hyperliens contenus par la page en question et enfin les ajoute à la liste des URL de sorte qu'il puisse enregistrer toutes les pages d'un site internet. Toutefois, l'enregistrement via crawling du contenu dynamique de bases de données ou de médias sociaux par exemple constitue un défi majeur. En revanche, le crawling

a l'avantage de pouvoir enregistrer de grands nombres de sites web. Une autre approche s'impose cependant pour les médias sociaux, à savoir l'extraction (harvesting) via des interfaces spécifiques ou API (Application Programming Interface) permettant d'extraire des données d'un système pour les mettre à disposition via un autre système. L'API de Twitter par exemple permet d'exporter toutes les activités Twitter d'un profil donné.

L'archivage web à l'étranger

Il existe à l'étranger de nombreuses initiatives d'archivage du web, dont certaines ont déjà été lancées à la fin des années 1990. Pour pouvoir tirer un enseignement de ces expériences, la littérature en la matière a été étudiée en profondeur. Ensuite, quelques initiatives d'archivage national du web à l'étranger ont été sélectionnées pour les analyser plus en détail. Les pays concernés sont le Royaume-Uni, la France, les Pays-Bas, l'Irlande, le Danemark, le Portugal, le Canada, le Luxembourg et la Suisse. Des représentants des établissements en question ont été interviewés et il a été procédé à une analyse détaillée du cadre légal au sein duquel l'archivage web est effectué ainsi que de la façon dont ces établissements sélectionnent, enregistrent et mettent à la disposition du public les sites à archiver.

Image 1: Les initiatives étrangères d'archivage du web qui ont été sélectionnées

Résultats de la recherche

À ce stade de la recherche, les résultats ont fait apparaître que dans ces pays l'archivage du web est effectué plutôt par la Bibliothèque nationale au lieu des Archives nationales, tandis que dans certains pays, les deux établissements constituent leurs propres archives du web.

Une grande variation peut être notée en ce qui concerne la politique de sélection. Au sein des bibliothèques nationales, deux grandes tendances peuvent être distinguées, à savoir le crawling large et le crawling sélectif. L'approche large vise tout le web national et en enregistre annuellement les niveaux supérieurs (c'est-à-dire la page d'accueil et les deux niveaux suivants par ex.). Pour ce faire, de nombreuses bibliothèques nationales collaborent avec les organismes responsables de la gestion des noms de domaine nationaux.

Les crawls sélectifs, quant à eux, constituent des collections collectées sur des thèmes spécifiques (Brexit, Harry Potter, …), des événements (élections, festivals, …) ou des situations inattendues

(calamités naturelles, attentats, …). Les sites web traités via un crawl sélectif sont archivés plus fréquemment et de façon plus poussée que ceux qui sont enregistrés via un crawl plus large.

Certaines bibliothèques nationales se limitent à un crawling sélectif, tandis que d'autres combinent les deux méthodes. Les archives nationales, quant à elles, s'occupent essentiellement de la collection de sites des pouvoirs publics. Une grande variation se présente également en ce qui concerne le traitement des médias sociaux : ces médias ne sont pas archivés par tous les établissements analysés, notamment parce que leur contenu est difficile à enregistrer. L'approche varie d'un réseau social à un autre, mais les plus populaires à être archivés sont Twitter, Facebook et YouTube.

La plupart des projets que nous avons analysés gardent un contrôle technique complet du processus d'archivage, mais il existe aussi des organisations qui confient la gestion entière à un prestataire de service externe. Le format de fichier le plus fréquemment utilisé pour des sites web archivés est le WARC (ISO 28500). Les fichiers WARC peuvent être mis à la disposition du public via un logiciel de « retransmission » faisant fonction de serveur web et affichant le contenu dans un navigateur internet.

Les modalités d'accès sont également très diverses : certaines archives du web ne sont pas du tout accessibles ou réservées aux chercheurs tandis que d'autres sont librement consultables en ligne. Le principal élément limitant l'accès est le droit d'auteur sur le contenu intellectuel. Généralement, les collections d'archives du web conservées par les bibliothèques nationales peuvent uniquement être consultées intra-muros au départ d'ordinateurs spécifiques où certaines fonctionnalités comme l'impression, la copie ou la capture d'écrans sont désactivées. D'autres bibliothèques nationales permettent un libre accès en ligne à une partie de leurs collections, à savoir aux sites web pour lesquels les titulaires du droit d'auteur ont permis de les mettre à disposition, ce qui implique toutefois une lourde administration pour en arriver là. Ce serait donc une erreur de croire que les sites web archivés seraient tout aussi accessibles que leurs contreparties actives. Généralement, les archives web conservées par les services nationaux d'archives sont (en partie) librement accessibles en ligne, étant donné que la majeure partie du contenu archivé provient de sites internet ou de médias sociaux d'organismes publics, et donc tombant sous le coup de la loi sur les archives.

À noter également que la recherche dans des archives web n'est pas aussi aisée que sur le web « en direct ». Toutes les archives web ne disposent pas d'une option de recherche en texte intégral (*full text*) : il faut alors chercher via une URL spécifique, ce qui veut dire que l'utilisateur doit connaître au préalable cette adresse URL ou une partie de son contenu pour pouvoir mener une recherche.

Le cadre légal

En Belgique, l'archivage du web fait partie des missions légales des AGR et de KBR. Les missions de KBR ont été fixées dans l'arrêté royal du 19 juin 1837 portant constitution en établissement scientifique de la Bibliothèque royale de Belgique (modifié le 25 décembre 2016). Une des missions de KBR est la rédaction d'inventaires des sites web ayant un rapport avec ses missions. Le projet examine aussi quelles recommandations pourraient être émises pour intégrer dans la législation sur le dépôt légal également l'archivage du web via extraction (harvesting).

Le cadre légal de l'archivage d'institutions publiques est régi par la loi sur les archives du 24 juin 1955, modifiée par la loi du 6 mai 2009. Les missions des AGR sont quant à elles fixées dans un arrêté royal du 3 décembre 2009 et dans d'autres arrêtés royaux de 2010 notamment ceux sur la surveillance archivistique et le transfert des archives. Les AGR sont notamment en charge de la surveillance des archives des pouvoirs publics, indépendamment de leur support, et elles acquièrent, conservent et, le cas échéant, éliminent des archives publiques. Il importe donc de bien définir ce que sont les archives. L'arrêté royal du 18 août 2010 portant exécution de la loi sur les archives du 24 juin 1955 définit les

archives comme « tous les documents qui, quels que soient leur date, leur forme matérielle, leur stade d'élaboration ou leur support, sont destinés, par leur nature, à être conservés par une autorité publique ou par une personne privée, une société ou une association de droit privé, dans la mesure où ces documents ont été reçus ou produits dans l'exercice de leurs activités, de leurs fonctions ou pour maintenir leurs droits et obligations. » Les sites web et les médias sociaux peuvent donc être considérés comme étant des archives.

Vers une stratégie belge

Pendant la deuxième phase du projet, les conclusions qui seront tirées du premier stade seront mises en pratique dans le contexte belge. Au niveau opérationnel, une stratégie commune des AGR et de KBR est pour l'instant mise sur pied. Elle comprend notamment l'élaboration d'une politique de sélection et d'enregistrement d'informations en ligne, un contrôle de qualité, le traitement de métadonnées et de documentation, le stockage, la préservation et l'accès. Un des défis majeurs du projet a trait à la complexité des sites internet. Ceux-ci peuvent être construits avec du texte, du matériel audiovisuel, du contenu dynamique, etc. ce qui complique l'enregistrement de ces éléments. Un autre défi consiste à définir ce qu'est le « web belge ». En effet, sur la « Toile », il n'y a pas de circonscriptions territoriales bien définies et en se limitant aux noms de domaine nationaux, la totalité des sites web gérés par des Belges ne serait pas prise en compte.

L'avenir

Fin 2018-début 2019, les deux établissements entreront dans la phase du projet pilote, afin de dresser des listes d'URL qui seront soumises à un crawling. L'accès à ces archives pilotes du web sera testé et évalué. Dans une dernière phase, basée sur les résultats de l'analyse de ce projet pilote, des recommandations seront formulées pour implémenter un service d'archivage durable du web au niveau fédéral belge.

Références

Brügger, N. & Schroeder, R. (Eds.). (2017). *The web as history: Using web archives to understand the past and present.* London: UCL Press.

Milligan, I. (2016). Lost in the infinite archive: the promise and pitfalls of web archives. *International Journal of Humanities and Arts Computing, 10(1)*, 78-94. doi: 10.3366/ijhac.2016.0161.

Vlassenroot, E., Chambers, S., Di Pretoro, E. , et al. (2018). Web archives as a data resource for digital scholars. *Digital Scholar, 1(1)*. (forthcoming)

[cadres]

Le projet PROMISE

Lancé le 1 juin 2017 et financé par BELSPO dans le cadre du programme BRAIN.BE, le projet de recherche PROMISE (Preserving Online Multiple Information: towards a Belgian StratEgy) vise à élaborer une stratégie fédérale pour la conservation du web belge. À cet effet, KBR (promotrices : Sophie Vandepontseele et Nadège Isbergue) et les AGR (promoteurs : Rolande Depoortere et Sébastien Soyez) collaborent avec les universités de Gand (Research Group for Media, Innovation and Communication Technologies ; Ghent Centre for Digital Humanities) et Namur (Centre de Recherche Droit Sociétés) et avec la Haute École Bruxelles-Brabant (Unité de Recherche et de Formation en Sciences de l'Information et de la Documentation).

Pour de plus amples informations sur le projet : https://www.kbr.be/nl/project-promise et https://www.arch.be.

L'auteur

Friedel Geeraert est chercheuse auprès de la Bibliothèque royale de Belgique et des Archives de l'État. Elle travaille sur le projet PROMISE relatif à l'archivage du web en Belgique.

Images à insérer:

Bibliothèque royale Koninklijke Bibliotheek On-line

Bibliothèque royale de Belgique

Koninklijke Bibliotheek van België

Français
Nederlands

BIBLIOTHÈQUE ROYALE DE BELGIQUE
KONINKLIJKE BIBLIOTHEEK VAN BELGIË

KBR·2

Nederlands Français

Koninklijke Bibliotheek van België
Bibliothèque royale de Belgique

Kies uw taal
Nederlands

Choisissez votre langue
Français

Choose your language
English

Het Rijksarchief in België - Archives de l'État en Belgique
Staatsarchive in Belgien - State Archives in Belgium

Nederlands | Français | English

[Het Rijksarchief in België]

[Archives de l'État en Belgique]

[State Archives in Belgium]

Federaal Wetenschapsbeleid
Politique scientifique fédérale
Föderalen Wissenschaftspolitik
Belgian Science Policy

Het Rijksarchief in België

Archives de l'État en Belgique

State Archives in Belgium

Nieuw
Bezoekersreglement
RICHTLIJN BETREFFENDE HET ZELF UITVOEREN VAN FOTOGRAFISCHE OPNAMEN

Neu
Besucherordnung
RICHTLINIE ÜBER DIE PERSÖNLICHE ANFERTIGUNG VON FOTOGRAFISCHEN REPRODUKTIONEN

Nouveau
Règlement pour les visiteurs
DIRECTIVE CONCERNANT LA RÉALISATION PERSONNELLE DE REPRODUCTIONS PHOTOGRAPHIQUES

- **Het Rijksarchief in België**
  Uw taalkeuze wordt onthouden voor uw volgende bezoeken.
- **Archives de l'État en Belgique**
  Votre choix de langue sera retenu pour vos prochaines visites.
- **Die Staatsarchive in Belgien**
  Ihre Wahle
- **State Archives in Belgium**
  The language preference you choose here will be retained for your next visits.

Archives de l'État en Belgique	Rijksarchief in België	Belgisches Staatsarchiv	State Archives of Belgium
En français	In het Nederlands	Deutsche Version	In English
Votre profil :	Uw profiel:	Ihr Profil:	Your profile:
Chercheur	Onderzoeker	Forscher	Researcher
Généalogiste	Genealoog	Genealoge	Genealogist
Fonctionnaire	Ambtenaar	Beamte	Civil Servant
Journaliste	Journalist	Journalist	Journalist
Enseignant	Leerkracht	Lehrkraft	Teacher
Notaire / Géomètre	Notaris / Landmeter	Notar / Landmesser	Notary / Surveyor

belspo .be

Légendes:

Images 2-6 : Évolution de la page d'accueil de la Bibliothèque royale (1998-2018)

Images 7-11: Évolution de la page d'accueil des Archives de l'État (1998-2018)

Notes finales:

Les versions historiques des pages d'accueil des Archives de l'État et de la Bibliothèque royale font partie des collections des archives du web.

Nog toe te voegen: screenshot van nieuwe homepage van KBR (lancering voorzien eind november)