PROMISE : UN PROJET DE RECHERCHE POUR UN ARCHIVAGE DU WEB BELGE AU NIVEAU FÉDÉRAL

Rolande DEPOORTERE
Chef du service Archivage digital, AGR
Friedel GEERAERT
Assistante scientifique en archivage du web, KBR
Gerald HAESENDONCK
IDLab UGent
Sébastien SOYEZ
Chef de travaux Service Archivage digital, AGR
Sophie VANDEPONTSEELE
Directrice des Collections contemporaines, KBR

Het artikel is opgesteld naar aanleiding van een colloquim "Saving the web: the promise of a Belgian web archive"[1], georganiseerd in KBR op 18 oktober 2019 te Brussel.

Article rédigé suite au colloque "Saving the web: the promise of a Belgian web archive"[2], organisé à KBR le 18 octobre 2019, à Bruxelles.

■ De juillet 2017 à décembre 2019, la Bibliothèque royale de Belgique (KBR) a lancé, en partenariat avec les Archives de l'Etat (AGR) et plusieurs universités et hautes-écoles belges, le projet de recherche PROMISE en vue de définir une stratégie fédérale pour l'archivage du web belge. Le présent article retrace le parcours mené par l'équipe de recherche durant plus de deux ans : état de l'art, définition d'une stratégie, étude de plusieurs scénarios et de leurs coûts et infrastructure technique. Les résultats de ce projet de recherche, financé par la Politique scientifique belge (BELSPO) ont également été présentés en octobre 2019 lors du colloque "Saving the web".

■ Van juli 2017 tot december 2019 werkte de Koninklijke Bibliotheek van België (KBR) – in samenwerking met het Algemeen Rijksarchief (ARA) en meerdere Belgische universiteiten en hogescholen – aan het onderzoeksproject PROMISE met als doel een federale strategie uit te stippelen voor de archivering van het Belgische web. Dit artikel schetst het parcours dat het onderzoeksteam gedurende meer dan twee jaar heeft afgelegd: state of the art, een strategie bepalen, studie van meerdere scenario's evenals hun kosten en technische infrastructuur. De resultaten van dit onderzoeksproject, dat werd gefinancierd door het Belgische wetenschapsbeleid (BELSPO), werden in oktober 2019 voorgesteld tijdens het colloquium "Saving the web".

Introduction au projet PROMISE

Le web occupe une place centrale dans la société et contient donc de nombreuses traces de notre histoire contemporaine. De ce fait, le web devient une ressource très intéressante pour les générations futures, raison pour laquelle différentes bibliothèques et archives (nationales) du monde entier archivent et préservent depuis des années, voire dans certains cas depuis des décennies, leur web national ou des parties de celui-ci, et donnent accès à ces collections.

En 2017 fut lancé au sein de KBR (Bibliothèque royale de Belgique) et des Archives de l'État (AGR), le projet de recherche *PROMISE*. Celui-ci avait pour objectif d'élaborer une stratégie fédérale de préservation du contenu du web belge. Le projet a été financé par la Politique scientifique belge (BELSPO) dans le cadre de leur programme BRAIN.be. En raison d'aspects techniques, juridiques et opérationnels et pour étudier les besoins des utilisateurs en matière d'archivage du web, les Archives de l'État et KBR ont collaboré avec les universités de Gand (Research Group for Media, Innovation and Communication Technologies; Ghent Centre for Digital Humanities) et de Namur (Centre de Recherche Information, Droit et Société (CRIDS)) et avec la Haute-École Bruxelles-Brabant.

Le projet, qui s'est clôturé en décembre 2019, était construit autour de quatre objectifs : 1) analyser les bonnes pratiques en archivage du web, 2) élaborer une stratégie d'archivage du web belge, 3) tester l'archivage du web et donner accès aux collections et 4) formuler des recommandations pour l'implémentation d'un service d'archivage du web durable.

L'objectif de cet article est d'offrir un aperçu de la structure globale du projet et des principaux résultats de recherche. La première partie contient un aperçu des bonnes pratiques qui ont été identifiées. La deuxième partie présente la stratégie esquissée dans le cadre du projet et la troisième partie présente les différents scénarios d'archivage du web au sein des Archives

de l'État et de KBR ainsi que l'analyse des coûts y afférents. Dans la dernière partie, l'infrastructure technique utilisée est présentée succinctement.

État de l'art

Dans le cadre du projet *PROMISE*, un certain nombre de projets d'archivage du web en Belgique et à l'étranger ont été étudiés. Certaines institutions patrimoniales belges archivent depuis des années du contenu du web, notamment : Felixarchief Antwerpen, Universiteitsbibliotheek Gent, Liberaal Archief, AMSAB - Instituut voor Sociale Geschiedenis, ADVN (Archief voor Nationale Bewegingen), KADOC, het Letterenhuis, Archief Gent ou l'Université Catholique de Louvain. De nombreuses formes d'expertises sont donc présentes en Belgique. La principale différence entre les ambitions de KBR et les Archives de l'État d'une part, et ces initiatives d'archivage du web, concerne les critères de sélection. KBR et les Archives de l'État souhaitent archiver le web belge de la manière la plus large possible alors que les institutions susmentionnées se concentrent sur certains sites web, et parfois aussi sur des médias sociaux, ayant un lien direct avec les priorités de leurs collections. Ceci est logique, mais on est en présence d'une réelle différence d'échelle. Le résultat est que KBR et les Archives de l'État ne peuvent généralement pas s'appuyer sur la même infrastructure ou approche que ces institutions, leurs pratiques n'étant pas modulables, comme par exemple l'utilisation de fichiers .zip dans une structure de dossiers déterminée. Les institutions patrimoniales belges constituent toutefois d'importantes parties prenantes pour le futur archivage du web belge puisque l'échange d'expertise et de pratiques entre les différentes institutions peut contribuer à la création d'une communauté pour l'archivage du web en Belgique.

Les initiatives en matière d'archivage du web qui ont été étudiées à l'étranger ont été sélectionnées de manière spécifique afin d'obtenir un bon mélange d'initiatives qui ont beaucoup d'expérience en matière d'archivage du web, qui opèrent dans des pays de différente taille, avec plusieurs langues nationales ou dans lesquels aussi bien les archives nationales que la bibliothèque nationale s'occupent de l'archivage du web et d'initiatives qui gèrent tous les aspects de l'archivage du web en interne ou collaborent avec des prestataires de services externes. Les initiatives ont été examinées au moyen d'une étude bibliographique complétée d'interviews semi-structurées avec des représentants des institutions en question. Les institutions suivantes ont été étudiées : la Koninklijke Bibliotheek et le Nationaal Archief aux Pays-Bas, la Bibliothèque nationale de France et l'Institut national de l'audiovisuel en France, la Bibliothèque nationale de Luxembourg, la British Library et les UK National Archives en Grande-Bretagne, la Kongelige Bibliotek au Danemark, l'Arquivo.pt au Portugal, l'Irish National Library et les Bibliothèque et Archives Canada et la Bibliothèque et archives nationales de Québec.

La politique menée au sein de ces institutions a été comparée sur quatre niveaux : sélection, accès, contexte juridique et infrastructure technique. Voici la problématique de la sélection et l'accès.

En ce qui concerne la sélection, la plupart des bibliothèques nationales combinent, dans des pays qui disposent d'un dépôt légal, des "crawls"[3] larges avec des "crawls" sélectifs. Les "crawls" larges servent à archiver le web national de manière superficielle alors que les collections sélectives visent généralement certains thèmes, événements ou circonstances imprévues. Les sites web qui sont archivés dans des collections sélectives, le sont de manière plus approfondie que ce n'est le cas pour les "crawls" larges. Il est aussi important de noter qu'il n'existe pas de définition univoque de ce que couvre exactement un web national et que les pratiques diffèrent donc d'un pays à l'autre. Quant aux archives nationales, elles constituent des collections de sites web d'institutions dont les archives doivent être déposées chez elles selon la loi sur les archives.

La moitié des institutions étudiées constitue, outre les sites web, des collections de médias sociaux. *Twitter* est archivé le plus fréquemment, suivi de *YouTube*, *Facebook*, Instagram et *Flickr*. Les institutions en question précisent toutefois que chaque canal de médias sociaux requiert une approche spécifique et que les API[4] sous-jacentes et l'infrastructure technique de certaines plateformes changent tellement vite qu'il faut pratiquement un suivi journalier pour permettre l'archivage de ce contenu de manière permanente. En outre, certaines plateformes permettent à peine d'être archivées. Initialement, l'équipe du projet *PROMISE* avait souhaité tester l'archivage aussi bien des médias sociaux que des sites web. Sur base des résultats des interviews, il a toutefois été décidé d'exclure l'archivage des médias sociaux du projet pour des raisons de faisabilité.

La manière de donner accès aux collections diffère aussi très fortement. En raison de la législation sur les droits d'auteur, la plupart des archives web sont uniquement accessibles dans les salles de lectures des institutions sauf si l'institution a reçu l'autorisation explicite des ayant-droits de donner accès librement au contenu du web archivé. Cette contrainte décourage en toute logique l'utilisation des archives du web étant donné que le public est habitué au web "live" qui est généralement en accès libre. Faire le pas vers la salle de lecture d'une institution est pour beaucoup un pont trop loin dans ce contexte. L'accès

à certaines archives du web comme au Danemark est encore plus restreint, notamment à un certain groupe d'utilisateurs, à savoir les chercheurs.

Non seulement l'endroit où des archives du web peuvent être consultées est important, mais aussi la manière dont les collections seront mises à disposition. La recherche par URL est possible dans toutes les institutions étudiées, mais la recherche plein texte ne l'est pas toujours. Cela est principalement dû à la taille des collections et donc aussi aux index établis sur la base du texte présent. Il n'est pas possible de créer l'infrastructure nécessaire pour toutes les initiatives. Dans de rares cas, il est également possible de parcourir les collections alphabétiquement ou thématiquement, mais cela ne concerne que certaines petites collections d'archives du web[5].

L'étude d'autres initiatives d'archivage du web était une phase très intéressante du projet. En effet, l'équipe du projet *PROMISE* a permis de tirer des leçons importantes des expériences acquises par d'autres institutions.

Stratégie

Pour développer la stratégie, KBR et les Archives de l'État (AGR) ont tenu compte des résultats de l'état de l'art, de l'analyse juridique réalisée par l'Université de Namur ainsi que des résultats d'une enquête de l'Université de Gand sur les besoins des utilisateurs dans le contexte de l'archivage du web.

Cette stratégie a été construite sur base du modèle d'exigences fonctionnelles de l'archivage électronique, à savoir celui publié dans la norme ISO 14721.2012, plus connu sous le nom d'OAIS[6]. Outre les 7 fonctions classiques que prévoit ce modèle (versement, gestion de données, stockage, planification de la préservation, accès, gestion opérationnelle et gestion stratégique), nous lui avons adjoint trois fonctions complémentaires, en phase avec les processus d'archivage du web, à savoir : la sélection, la collecte et le contrôle-qualité. Ce découpage fonctionnel nous a permis de construire notre stratégie sur base d'une structure logique, de la collecte jusqu'à la diffusion, en passant par la mise en archive.

Au niveau de la description du corpus de données, notre choix s'est porté sur un schéma existant qui permettait d'utiliser, tant pour KBR que pour les AGR, une liste de métadonnées identique. Il s'agit du schéma de métadonnées établi par le *Web Archiving Metadata Working Group* de l'OCLC[7]. Ce schéma, construit pour établir une liste de 14 descripteurs pour toute archive web[8], permet en outre de faire converger deux préoccupations professionnelles distinctes, à savoir celle du monde des bibliothèques et celle des archives. En effet, ce schéma peut aisément renvoyer vers les descripteurs de chaque domaine, à savoir MARC21[9] pour les bibliothécaires, et la DTD-EAD[10] pour les archivistes.

Dès que ces deux premiers choix stratégiques - le schéma fonctionnel et le schéma de métadonnées - ont été posés, il était important de définir quelle serait la base sélective de la collecte des archives web en Belgique. En effet, plusieurs choix pouvaient s'opérer. Première possibilité, celle de sélectionner l'ensemble des sites ayant ".be" comme noms de domaine génériques[11], éventuellement en y ajoutant les sites régionaux ou locaux[12]. Sur base des derniers chiffres connus[13], on peut estimer ce volume total à environ 1 million. Seconde possibilité, on peut définir des listes sélectives sur base de choix précis, comme par exemple la pertinence de sites web pour nos deux institutions. Pour établir ces listes, KBR et les AGR ont, par exemple, sélectionné des sites - ou des parties de sites - qui étaient particulièrement intéressants compte tenu de leurs missions spécifiques, ou qui constituaient une obligation légale de conservation. À titre d'exemple, pour KBR, des sites en lien notamment avec la littérature, la musicologie ou la BD, et pour les AGR, des sites (para-)publics liés au fonctionnement de l'État fédéral, régional ou local. Après analyse, KBR a retenu un peu moins de 1.000 sites et environ 1.400 pages. Les AGR quant à elles ont établi leurs listes sélectives selon 3 niveaux : environ 650 sites fédéraux, 1.400 sites régionaux et locaux, et 300 d'origine privée. Lors de la mise en place de la stratégie globale, il restera à choisir ou à combiner l'une ou l'autre de ces listes, et d'envisager le cas échéant la sélection d'un échantillon aléatoire[14].

La stratégie de la collecte proactive est une des options retenues par le projet[15]. Une option alternative pourrait être celle d'attendre que des gestionnaires de sites web nous versent, soit volontairement soit par obligation légale, leurs contenus informationnels. Mais la pratique nous démontre que les résultats pourraient être relativement peu uniformes et parcellaires. Outre ce choix, nos deux institutions devront également s'accorder sur la fréquence (annuelle) et sur la profondeur (complète ou quelques niveaux) de cette collecte. Pour effectuer techniquement cette collecte proactive, notre choix s'est porté sur l'outil *Heritrix*[16]. Son fonctionnement est relativement simple : lors de son passage sur une page web, cet outil prend une copie complète du contenu d'un site et le sauvegarde dans le format de fichier WARC[17]. Ce sera à partir de ce fichier WARC qu'une nouvelle consultation du site sera possible lors de l'étape de l'accès. Après cette collecte, il est indispensable d'effectuer un contrôle-qualité, soit systématique pour un volume limité de sites, soit sur base d'échantillons. Il existe deux possibilités

pour effectuer ce test de qualité : manuellement ou semi-automatiquement. Manuellement tout d'abord, en prenant un nombre limité de sites collectés et en les comparant avec la version d'origine, de manière visuelle mais également en testant les hyperliens. Cette méthode, très gourmande en ressources, est la seule qui permet de faire un contrôle systématique du contenu archivé. Sur base de notre expérience, il faut compter en moyenne 10 jours de travail pour contrôler environ 600 sites. Par ailleurs, il est possible de mettre en place des contrôles semi-automatisés sur des sites, sur base de paramètres techniques prédéfinis[18]. Dans le cadre du projet *PROMISE*, plusieurs paramètres ont été testés, mais ne constituent pas à eux seuls la solution. Notre conclusion porterait davantage sur une solution mixte, de contrôle manuel d'une partie limitée de la collection, et de contrôle semi-automatique des collectes larges.

Dès que la collecte et le contrôle-qualité ont été réalisés, les collections sont ensuite transférées dans les dépôts de chaque institution, par le biais de leur procédure de transfert en place. Ensuite, il est indispensable de régler la question de la mise en archive proprement dite, à savoir de mettre en place une gestion des (méta)données, une infrastructure de stockage ainsi que garantir la pérennisation des objets numériques archivés. Pour ce qui est de la gestion de (méta)données, chaque institution (KBR et AGR) peut choisir son mode de gestion propre, pour autant que la description avec le schéma commun soit maintenue. Les données administratives de gestion seront intégrées dans chaque catalogue spécifique (*Syracuse* pour KBR et *SAM* pour les AGR), et il sera décidé de leur adjoindre les métadonnées de description complémentaires, à savoir MARC21 pour KBR, et la DTD EAD pour les AGR. Cette transposition sera facilitée par le choix initial du schéma de métadonnées de l'OCLC. La liaison entre ces métadonnées et les fichiers WARC, constituant l'archive web, sera effectuée à l'aide de fichiers METS. Pour ce qui est du stockage, deux stratégies sont envisageables : soit chaque institution gère sa propre collection sur ses propres espaces de stockage, soit elles mutualisent une de leurs *infra*structures communes prévue par le LTP[19], financée actuellement par la Politique scientifique fédérale (BELSPO). Ces choix devront principalement conduire à une rationalisation des coûts de stockage (cf. *infra*). Les deux institutions avertiront la plateforme LTP lorsqu'un format déterminé sera arrivé en fin de vie. Chaque année, tous les formats de fichiers se trouvant dans les collections de l'archive web seront identifiés sur base de quoi un tableau de contrôle sera établi. Les institutions effectueront un monitoring technologique continu pour rester au courant des derniers développements.

Enfin, plusieurs éléments concernant la stratégie de l'accès se sont dégagés au terme du projet *PROMISE*. Tout d'abord, d'un point de vue de l'accès physique, et en respectant l'actuelle législation sur le droit d'auteur, il ne sera possible d'accéder à certains sites web que par le biais de l'une des salles de lecture de nos institutions. Bien entendu, cette restriction ne s'appliquera qu'aux sites web soumis à ce droit d'auteur. Pour la majorité des sites web créés par des autorités publiques, ce problème ne se posera pas, et une consultation directement en ligne est envisagée. En ce qui concerne l'interface d'accès aux collections web de KBR et des AGR (cf. *infra*), construite autour du modèle commun de description (OCLC), il est envisagé de développer les fonctionnalités suivantes : recherche par URL, recherche plein texte et par mots-clés et recherche par organismes "responsables" du site. L'interface se basera sur un "replay" des sites[20], à partir des collections archivées. Il est également envisagé de créer une interface de récupération de sets de données plus larges, ce qui permettrait aux chercheurs/ses d'analyser avec d'autres outils ces contenus de données.

Scénarios et calcul des coûts

Il est difficile de présenter tous les scénarios envisagés tant les combinaisons possibles sont nombreuses. En effet, si l'on se base sur les différents types de collecte, dont les coûts sont mutualisés entre KBR et AGR, ou non, et ce pour chaque phase du modèle OAIS, il existe plus de cent combinaisons possibles. Il a été nécessaire de faire un choix et quatre scénarios combinant un intérêt tant sur l'approche de la sélection que sur la mutualisation des coûts ont été sélectionnés.

Dans l'objectif de prendre une décision quant à la définition d'une politique structurelle pour l'archivage du web belge, la KBR et les AGR ont travaillé sur différents scénarios d'archivage du web. Cette analyse a permis d'offrir une variété d'approches institutionnelles possibles de l'archivage web en fonction des ressources qui peuvent être mises à disposition. Ces scénarios sont basés sur différentes approches concernant la sélection et couvrent trois niveaux différents : complet, intermédiaire et basique.

Le scénario complet couvre la collecte de collections sélectives et un large éventail comprenant la collecte de 100% du web belge. En ce qui concerne les AGR, les collections sélectives comprendraient d'abord les sites web des institutions fédérales dont les archives doivent être légalement conservées, ensuite les sites web des villes, des communes et des organismes parapublics et enfin les sites web des archives privées. Pour KBR, les collections sélectives

comprendraient des sites web étroitement liés aux collections existantes et s'inscrivant pleinement dans la charte de développement des collections qui définit les grands principes d'acquisition des collections. Les sites web qui font partie des collections sélectives seraient collectés dans leur entièreté. Par contre, dans le cas de la collecte large, la sélection se limiterait à parcourir uniquement les couches supérieures des sites web, constituant ainsi un échantillon.

Dans le scénario intermédiaire, la collecte large serait limitée à un échantillon choisi au hasard à hauteur de 10 % du web belge. Les collectes sélectives pour les AGR seraient limitées aux sites web des institutions fédérales soumises à la loi sur les archives. Pour KBR, les collections sélectives couvriraient le même contenu que dans le scénario complet.

Le scénario dit basique comprendrait les mêmes collections sélectives que dans le scénario intermédiaire, mais en excluant la collecte large.

Un quatrième scénario a également été envisagé : il s'agit de celui de l' " outsourcing ". Cette piste pourrait constituer une alternative intéressante dans le cas où il ne serait pas possible d'effectuer l'archivage du web belge en interne. Il semble aussi intéressant de connaître le prix de ce type de prestation et de le comparer au scénario le plus complet. Cependant, cette piste présente quelques limites : si l'abonnement est stoppé, il sera alors nécessaire de développer une infrastructure propre pour consulter les fichiers WARC.

Afin d'estimer le coût annuel de l'archivage du web belge, nous avons listé toutes les tâches à exécuter à chaque étape du modèle OAIS, en distinguant les tâches à répéter chaque année, les coûts annuels de maintenance et les investissements périodiques pour garantir la pérennité de l'infrastructure. Le calcul a porté sur les ressources humaines à mobiliser et sur l'infrastructure douce et dure, incluant le coût de la maintenance et des futures mises à jour. La projection porte sur une période de cinq ans, dont a été tirée une moyenne annuelle. Les ressources humaines ont été estimées en nombre d'heures pour effectuer les traitements, sur base de tests effectués sur un échantillon de sites web capturés. Le calcul du coût humain intègre le salaire horaire moyen des différents profils de compétences nécessaires, en fonction de leur niveau de qualification, pour trois familles de compétences identifiées: "archiviste numérique", "bibliothécaire numérique" et "informaticien". Par scénario envisagé, le total général est exprimé en équivalent temps plein (ETP). Il été tenu compte du salaire élevé des métiers en pénurie.

Dans le scénario le plus complet, la collecte sélective couvre l'intégralité de 2.350 sites pour les AGR et de 920 sites pour la KBR, plus 1.400 pages d'autres sites pour la KBR. S'ajoute à cette sélection l'échantillon du million de sites web belges moissonnés chacun partiellement. Ce scénario exige 5,76 ETP en personnel, soit un montant de 275.000 euros. La mise à jour de la liste des sites à moissonner intégralement, le processus de collecte (incluant le contrôle qualité de ces sites) et la gestion opérationnelle sont les tâches les plus gourmandes en personnel, la collecte large du million de sites étant entièrement automatisée sans contrôle qualité humain. Le volume de la collecte est estimé à 114 Tb pour la première année, dont 40 Tb pour la collecte large du million de sites et 70 Tb pour la collecte sélective des sites intéressant la KBR. Ce volume considérable s'explique par la nature et la fréquence de la collecte : certains sites de presse se composent de beaucoup de fichiers lourds (vidéos, sons, images fixes) et ils seront moissonnés plusieurs fois par jour en raison de leur caractère ultra-dynamique. En comparaison, la collecte sélective des 2.350 sites intéressant les Archives de l'État n'atteint que 4 Tb parce que ces sites sont moins souvent modifiés et qu'il a été jugé suffisant de les moissonner une fois par an, en complément à la collecte d'autres archives numériques des administrations concernées. Un accroissement annuel de 10% a été intégré dans le calcul. L'infrastructure coûterait environ 205.000 euros. Au total, le scénario le plus complet revient à 480.000 euros annuels.

Dans le scénario intermédiaire, le nombre de sites à moissonner intégralement est fort réduit pour les Archives de l'État mais reste identique pour KBR, tandis que la collecte large est divisée par dix. Le volume total annuel de stockage descend à 75 Tb, le coût global de l'infrastructure baisse à 136.500 euros, soit une réduction due uniquement à l'impact sur la capacité de stockage, car le reste de l'infrastructure douce et dure est le même que dans le scénario complet. Les ressources humaines ne diminuent que d'1 ETP car la collecte large se fait sans intervention humaine. Le coût total de ce scénario est estimé à 360.000 euros.

Dans un troisième scénario dit basique, la collecte sélective est identique à celle du scénario intermédiaire mais aucun site supplémentaire n'est moissonné. Or ceci n'entraîne pas de réduction notable des coûts. Les besoins en personnel restent identiques puisque, la collecte large n'impliquant pas d'intervention humaine, son abandon n'impacte pas les ressources à prévoir. La capacité de stockage est réduite de 4 Tb. Au total, l'économie entre le scénario intermédiaire et le basique se limite à 6.700 euros.

Infrastructure technique

Un des objectifs du projet *PROMISE* était de tester l'archivage du web et de donner accès aux collections. Pour ce faire, il fallait bien sûr également prévoir l'infrastructure technique nécessaire, surtout en ce qui concerne la sélection, la collecte et l'accès.

Un prototype du module de sélection a été développé. Ce module permet d'introduire un URL et de créer manuellement les métadonnées descriptives nécessaires basées sur le modèle OCLC pour métadonnées descriptives dans les archives du web[21]. Ce module, développé par la Haute-École Bruxelles-Brabant, est basé sur Python, Django et PostgreSQL. Le logiciel pourrait être développé davantage dans le futur afin de pouvoir commander l'exploration et le contrôle qualité.

Afin de capturer les sites web, le logiciel *Heritrix* a été utilisé dans le cadre du projet *PROMISE*. Ce "software", appelé également "web crawler", part d'une liste d'URLs sélectionnés[22]. En suivant les liens internes entre les différentes pages web, le "crawler" stocke une copie de tout le contenu des pages web et de leurs métadonnées techniques dans un fichier WARC[23]. Le format de fichier WARC est comparable à un container pour tous les contenus web déterminés et informations contextuelles afférentes. Il s'agit du format de fichier le plus courant pour le contenu du web archivé au niveau international. Le principal avantage de *Heritrix* est la rapidité de l'exploration et le fait que le logiciel, utilisé depuis longtemps, a déjà fait ses preuves. En revanche, ce logiciel a du mal à moissonner des sites web complexes, comme ceux avec de nombreux médias sociaux ou Javascript. Un certain nombre de tests ont dès lors été effectués avec des outils tels que *Browsertrix* et *Brozzler*, spécialement développés pour capturer des sites web au contenu dynamique[24]. Ces outils donnent des résultats de grande qualité, mais utilisent nettement plus de puissance de calcul que *Heritrix*. En outre, ces outils sont récents et donc pas encore stables.

Le prototype du module d'accès est basé sur *WARCLight*, une application spécialement axée sur la découverte d'éléments dans une archive, et sur *pyWB*[25]. *PyWB* est le "replay software" qui permet aux utilisateurs d'interagir avec un site web archivé comme sur le "live web". *PyWB* permet d'afficher une version spécifique avec un certain horodatage d'un site web. Actuellement, le prototype permet uniquement des recherches sur base d'URL, mais l'objectif est de développer, dans le futur, davantage de possibilités de recherche et de filtres.

Conclusion

Le projet de recherche *PROMISE* constitue une étape fondamentale pour l'implémentation d'une politique structurelle d'archivage du web au niveau fédéral. L'étude des meilleures pratiques en Belgique et à l'étranger a permis de tirer des leçons des expériences des institutions qui s'occupent de l'archive du web depuis un certain temps. Ces leçons se trouvaient, entre autres, à la base de la stratégie commune qui a été développée dans le cadre du projet *PROMISE*. Actuellement, l'archive du web belge n'existe que comme prototype, mais le but est de développer une archive du web belge fonctionnelle dans les prochaines années. Les scénarios concrets et les coûts afférents peuvent servir de base à KBR et aux Archives de l'État pour prendre des décisions stratégiques ultérieures. Il faudra évidemment tenir compte des moyens disponibles, tant au niveau financier qu'au niveau des ressources humaines. Les conclusions de ce projet constituent, par ailleurs, une opportunité pour défendre le besoin de disposer d'un budget structurel dédié à la gestion, tant pour les questions de préservation que pour améliorer l'accès pour le citoyen, des archives numériques qu'elles proviennent de la numérisation ou qu'elles soient nativement numériques.

Rolande Depoortere
Sébastien Soyez
Archives Générales du Royaume
Rue du Ruysbroeck 2
1000 Bruxelles
rolande.depoortere@arch.be
sebastien.soyez@arch.be
http://www.arch.be
Friedel Geeraert
Sophie Vandepontseele
KBR
Boulevard de l'Empereur, 4
1000 Bruxelles
friedel.geeraert@kbr.be
sophie.vandepontseele@kbr.be
http://www.kbr.be
Gerald Haesendonck
IDLab UGent
Technologiepark-Zwijnaarde 126
9052 Gent
gerald.haesendonck@UGent.be
https://www.ugent.be/ea/idlab

Mai 2020

Références

Archives Unleashed. *Warclight* [en ligne]. <https://github.com/archivesunleashed/warclight> (consulté le 23 janvier 2020).

International Organisation for Standardisation. *ISO 28500:2017. Information and documentation — WARC file format* [en ligne]. 2017 (consulté le 23 janvier 2020) <https://www.iso.org/standard/68004.html>.

Dooley, Jackie & Bowers, Kate. *Descriptive metadata for web archiving. Recommendations of the OCLC Research Library Partnership Web Archiving Metadata Working Group.* Online Computer Library Center, 2018 (consulté le 23 janvier). Technical report. <https://www.oclc.org/research/publications/2018/oclcresearch-descriptive-metadata.html>.

Internet Archive. *Heritrix* [en ligne]. <https://github.com/internetarchive/heritrix3/wiki> (consulté le 23 janvier 2020).

Internet Archive. *Brozzler* [en ligne]. <https://github.com/internetarchive/brozzler> (consulté le 23 janvier 2020).

Vlassenroot, Eveline, Chambers, Sally, Di Pretoro, Emmanuel, Geeraert, Friedel, Haesendonck, Gerald, Michel, Alejandra, Mechant, Peter. Web archives as a data resource for digital scholars. *International Journal of Digital Humanities* [en ligne], mars 2019 (consulté le 23 janvier 2020), vol. 1, n°1. <https://link.springer.com/article/10.1007/s42803-019-00007-7>.

Webrecorder. *Browsertrix* [en ligne]. <https://github.com/webrecorder/browsertrix> (consulté le 23 janvier 2020).

Webrecorder. *PyWB* [en ligne]. <https://github.com/webrecorder/pywb> (consulté le 23 janvier 2020).

Notes

1. <https://www.kbr.be/nl/colloquium-saving-the-web-the-promise-of-a-belgian-web-archive/>
2. <https://www.kbr.be/fr/colloque-saving-the-web/>
3. To crawl = collecter, moissonner.
4. Application Programme Interface.
5. Vlassenroot, Eveline et al. Web archives as a data resource for digital scholars. *International Journal of Digital Humanities* [en ligne], mars 2019 (consulté le 23 janvier 2020), vol. 1, n°1. <https://link.springer.com/article/10.1007/s42803-019-00007-7>.
6. OAIS : pour *Open Archive Information System*, un Système Ouvert d'Archivage d'information <https://public.ccsds.org/Pubs/650x0m2%28F%29.pdf> (consulté le 30 mars 2020).
7. OCLC, pour *Online Computer Library Center* <https://www.oclc.org/research/publications/2018/oclcresearch-descriptive-metadata.html>
8. La description selon 14 métadonnées prend du temps et n'est concrètement possible que pour un nombre limité de sites web. En fonction des choix stratégiques de collecte sélective et/ou large, il est envisagé de limiter ces descriptions à 2 ou 3 métadonnées, obtenues de manière automatisée.
9. MARC 21, cf. <https://www.loc.gov/marc/bibliographic, consulté le 31/03/2020>.
10. DTD-EAD, cf. <https://www.loc.gov/ead/ead2002a.html, consulté le 31/03/2020>.
11. En anglais, on parle de "ccTLD" pour "Country Code Top Level Domain", <https://www.dnsbelgium.be/fr/nouvelles/les-extensions-de-noms-de-domaine-et-le-monde> (consulté le 31 mars 2020)
12. Comme par exemple pour les régionaux .brussels, ou .vlaanderen, et pour des locaux .gent.
13. Cf. les rapports annuels de DNS, <https://www.dnsbelgium.be/fr>,consulté le 31 mars 2020.
14. Cet échantillon pourrait par exemple comprendre de collecter une partie seulement de tous les sites ".be" (3-4 niveaux de profondeur), ou 10% de tous les sites ".be" de manière intégrale.
15. Cette stratégie pourrait évoluer suivant l'adaptation du cadre légal en Belgique, notamment lorsque le web sera intégré comme élément du dépôt légal numérique, ou quand le droit d'auteur tiendra compte de la particularité patrimoniale du web.
16. Cf. page web de l'outil : <http://crawler.archive.org/index.html>,consulté le 31 mars 2020.
17. WARC : Web ARChive format, est un format d'archivage de sites web reconnu par l'ISO : <https://www.iso.org/fr/standard/68004.html>, consulté le 31 mars 2020.
18. Ces paramètres peuvent être la correspondance visuelle, la correspondance interactive, la complétude, la pertinence de la taille et du contenu. Référence : Ayala Brenda Reyes, *A Grounded Theory of Information Quality in Web Archives* (PhD, 2018), cf. <https://digital.library.unt.edu/ark:/67531/metadc1248497>, consulté le 26 mai 2020.

19. LTP : *Long Term Preservation Platform*, cf. <https://www.belspo.be/belspo/organisation/doc/Org/Contrat%20d%27administration%202016-2018%20SPP%20PS>.pdf, consulté le 31 mars 2020.

20. Ce "replay" des sites web sera basé sur la technologie développée par *Internet Archive*, à savoir la *Wayback Machine*. Il peut être mis en œuvre techniquement en installant par ex. *PyWB* (Python Wayback Machine), cf. <https://github.com/webrecorder/pywb, consulté le 31/03/2020>.

21. Dooley, Jackie & Bowers, Kate. *Descriptive metadata for web archiving. Recommendations of the OCLC Research Library Partnership Web Archiving Metadata Working Group*. Online Computer Library Center, 2018 (consulté le 23 janvier). Technical report. <https://www.oclc.org/research/publications/2018/oclcresearch-descriptive-metadata.html>.

22. Internet Archive. *Heritrix* [en ligne]. <https://github.com/internetarchive/heritrix3/wiki> (consulté le 23 janvier 2020).

23. International Organisation for Standardisation. *ISO 28500:2017. Information and documentation — WARC file format* [en ligne]. 2017 (consulté le 23 janvier 2020) <https://www.iso.org/standard/68004.html>.

24. Webrecorder. *Browsertrix* [en ligne]. <https://github.com/webrecorder/browsertrix> (consulté le 23 janvier 2020). Internet Archive. *Brozzler* [en ligne]. <https://github.com/internetarchive/brozzler> (consulté le 23 janvier 2020).

25. Archives Unleashed. *Warclight* [en ligne]. <https://github.com/archivesunleashed/warclight> (consulté le 23 janvier 2020). Webrecorder. *PyWB* [en ligne]. <https://github.com/webrecorder/pywb> (consulté le 23 janvier 2020).