PROMISE : préserver les multiples informations en ligne : vers une stratégie belge

RÉSUMÉ

Contexte

L'internet, devenu un moyen de communication de référence de notre quotidien, représente une matière importante du point de vue patrimonial. En tant que collections de données ou de documents, les sites web constituent dès lors des matériaux essentiels à archiver. Aujourd'hui, le web est aussi considéré comme un moyen de publication à part entière. À l'instar d'autres publications et archives, dont la préservation est garantie par la législation sur le dépôt légal et la loi sur les archives, le web belge doit faire l'objet d'une politique de préservation à long terme.

Objectifs

Le projet PROMISE a été lancé afin de répondre à la question urgente de la préservation du web belge - en tant que partie importante de l'histoire belge - pour les générations futures. L'objectif du projet est de développer une stratégie fédérale de préservation du web belge. Ce projet se déploie en quatre étapes :

- Identifier les bonnes pratiques en matière d'archivage du web
- Définir une politique belge d'archivage du web au niveau fédéral
- Mettre en place un projet pilote d'archivage du web, de son accès et son utilisation pour l'étude scientifique de celui-ci
- Formuler des recommandations pour développer un service d'archivage du web durable pour la Belgique

Méthodologie

D'un point de vue méthodologique, le projet PROMISE a d'abord collecté des données (de textes légaux, d'initiatives existantes, du rôle des institutions) par le biais d'une analyse documentaire et d'interviews semi-structurées avec des représentants d'initiatives étrangères en matière d'archivage du web. Sur la base des informations collectées, une stratégie et une politique réalistes ont été développées afin de collecter et de préserver le contenu en ligne du web belge. Cette stratégie comprend un calcul élaboré des coûts basé sur des scénarios réalistes d'archivage du web. Une enquête a aussi été menée afin d'analyser les besoins des usagers en matière d'archives du web. Ses résultats ont été pris en compte dans l'élaboration de la stratégie.

Le modèle qui a été développé a ensuite été validé par l'établissement d'un archivage du web pilote, comprenant la sélection et l'extraction du contenu et l'ouverture des collections pour leur accès et leur utilisation. Et enfin, le projet PROMISE a formulé des recommandations pour implémenter un service d'archivage du web durable en Belgique incluant des perspectives légales et opérationnelles. Sur le plan

législatif, une étude approfondie a été menée concernant la législation sur la protection des données personnelles au niveau européen (GDPR) et belge. L'aspect opérationnel comprenait le développement d'un business model basé sur le Service-Dominant Business Model et l'établissement de procédures opérationnelles tenant compte des résultats de recherche obtenus dans le cadre du projet PROMISE.

Conclusions

Le projet PROMISE a établi un rapport détaillé sur l'état des lieux des bonnes pratiques internationales en matière d'archivage du web, tenant compte des aspects légaux, opérationnels et techniques. Les principales conclusions de cette phase du projet ont aussi été publiées sous la forme d'un article scientifique (Vlassenroot et al., 2019, voir https://link.springer.com/article/10.1007/s42803-019-00007-7).

En vue de définir une politique belge de l'archivage du web au niveau fédéral, un rapport a été établi concernant le cadre légal autour de l'information du web belge. Une note stratégique pour le Conseil des Directeurs des Archives de l'État et de KBR a également été rédigée. La note comprenait les différentes phases du processus d'archivage du web (sélection, capture, intégration, préservation, accès…) ainsi qu'une analyse des coûts détaillée pour un archivage fonctionnel du web basé sur des scénarios réalistes. Une enquête sur les besoins des utilisateurs en matière d'archives du web a aussi été menée. Ses résultats ont été publiés dans un dashboard interactif (voir https://public.tableau.com/profile/eveline.vlassenroot#!/vizhome/PReservingOnlineMultipleInformationtowardsaBelgianstrategy/PReservingOnlineMultipleINformationtowardsaBelgianstrategy).

La phase pilote du projet a consisté à sélectionner le contenu du web à archiver ainsi que des outils nécessaires, à extraire ce contenu et à piloter l'accès à ces collections. Des listes d'adresses (seed lists ou listes d'URL pertinents) ont été établies pour les Archives de l'État et KBR. Le contenu a été capturé moyennant l'outil "open source" Heritrix, ce qui a donné lieu à une collection de fichiers WARC. L'accès aux archives du web a été réalisé par le biais de l'outil "open source" PyWB dont une instance a été installée sur les serveurs des Archives de l'État et de KBR. Le contenu capturé, ainsi que les outils implémentés, ont fait l'objet d'une évaluation itérative et informelle durant la durée du projet. Il s'agissait d'un exercice d'évaluation de la qualité et de l'exhaustivité du matériel web archivé. Dans ce contexte, on a essayé de répondre à des questions telles que (i) « Quel est le pourcentage de l'histoire belge qui a été perdu du fait qu'il n'existe pas d'archivage du web belge ? », (ii) « Quels sites web semblent résister à l'épreuve du temps et sont toujours en ligne ? » et (iii) « Quelle partie de l'ancien web belge peut être reconstruite via d'autres archives du web ou en utilisant des techniques d'archéologie du web ? ». Les résultats (parmi d'autres) ont été présentés lors de la conférence 'Unearthing the Belgian web of the 1990's: a digitised reconstruction' qui a eu lieu dans le cadre de la 3e RESAW Conference 'The Web That Was: Archives, Traces and Reflections' (Amsterdam, 19-21 juin 2019). Pour d'autres activités d'évaluation, on a utilisé des personae créées dans le projet Corpus et décrites

dans “Le projet Corpus et ses publics potentiels : Une étude prospective sur les besoins et les attentes des futurs usagers” (cf. https://hal-bnf.archives-ouvertes.fr/hal-01739730/document). Ainsi, ces cinq personae, initialement désignées pour contribuer à « l’identification et à la définition des profils des usagers potentiels » (p. 38), ont été utilisées pour évaluer les méthodes actuelles d’accès aux archives du web et les outils pour le faire.

En matière de recommandations pour un archivage du web durable, un certain nombre d’actions ont été entreprises. Premièrement, plusieurs recommandations ont été formulées en vue de réviser et de modifier la législation belge en matière de dépôt légal. Deuxièmement, un rapport approfondi concernant les considérations légales en matière d’accès aux archives du web a été établi. Ensuite, une liste de questions fréquentes (FAQ), a été préparée, liste qui peut être utilisée par les Archives de l’État et KBR comme ligne directrice pour la protection des données personnelles dans le contexte de l’archivage du web. Troisièmement, des « arbres de décision » ont été créés pour veiller à l’application du droit d’auteur pour l’archivage du web au niveau de la sélection et de l’accès. Quatrièmement, une FAQ sur la protection des données personnelles a été dressée afin d’aider le personnel de KBR et des AGR à comprendre les défis liés à l’archivage du web. Un business model a été développé pour KBR et les Archives de l’État et des procédures opérationnelles couvrant l’entièreté du flux de travail de l’archivage du web ont aussi été définies.

Mots-clés

- archivage du web
- humanités numériques
- préservation numérique
- information en ligne
- internet