PROMISE: Preserving Online Multiple Information: towards a Belgian strategy
SUMMARY

Context

The web has become a central means of communication in our everyday lives, which makes it very valuable from a heritage perspective. Websites, as collections of data and documents, are therefore important materials to be archived. Today the web is also considered as a publication channel in its own right. As is the case for other publications and archives, the preservation of which is guaranteed by legal deposit legislation and the law on archives, a long-term preservation policy needs to be developed for the Belgian web.

Objectives

The PROMISE project was initiated to formulate an answer to the urgent question of how to address the preservation of the Belgian web for future generations, as an important part of Belgian history. Four goals were addressed within the project:

- To identify current best practices in web archiving
- To define a Belgian policy for web archiving on the federal level
- To pilot web archiving, access and use of the pilot Belgian web archive for scientific research
- To make recommendations for a sustainable web archiving service for Belgium

Methodology

From a methodological point of view, the PROMISE project first collected data (from legal texts, existing initiatives, institutions' roles) by means of a literature review and in-depth semi-structured interviews with representatives of web archiving initiatives abroad. Based on the gathered information, a viable strategy and policy to capture and preserve online content on the Belgian web was drafted. This included an elaborate cost calculation based on realistic web archiving scenarios. A survey was also undertaken to analyse user requirements in web archives, the results of which were also taken into account in the strategy.

The model that was outlined was then validated by undertaking a pilot web-archive comprising of selecting and harvesting the content and opening up the collections for access and use. Lastly, the PROMISE project drafted recommendations for the implementation of a sustainable web archiving service in Belgium including legal and operational perspectives. For the former an in-depth study of personal data protection legislation on the European (GDPR) and Belgian level was undertaken. The latter comprised of developing a business model based on the Service-Dominant Business Model and defining operational procedures taking into account the research results obtained over the course of the PROMISE project.

Conclusions

The PROMISE project produced a detailed report about the state of the art of web archiving best practices internationally taking into account legal, operational and technical aspects. The main findings of this phase of the project were also published as a research article (Vlassenroot et al., 2019, see https://link.springer.com/article/10.1007/s42803-019-00007-7).

For the definition of a Belgian policy for web archiving at the federal level, a report on the legal framework surrounding Belgian web information was produced. A strategic note for the Board of Directors of the State Archives and KBR was also developed. The note encompassed the different phases in the web archiving process (selection, capture, ingest, preservation, access, …) and also included a detailed cost analysis for a functional web archive based on realistic scenarios. A survey about user requirements in the context of web archives was also conducted, the results of which were published as an interactive dashboard (see https://public.tableau.com/profile/eveline.vlassenroot#!/vizhome/PReservingOnlineMultipleInformationtowardsaBelgianstrategy/PReservingOnlineMultipleINformationtowardsaBelgianstrategy).

The pilot phase of the project consisted of selecting web content to be archived as well as the tools needed, harvesting this content and piloting access to these collections. Seed lists (i.e. lists of pertinent URLs) were created for the State Archives and KBR. The content was captured with the open source tool Heritrix, which resulted in a collection of WARC files. Access to the web archive was realised by means of the open source PyWB tool of which an instance was installed on the servers of the State Archives and KBR. The captured content, as well as the implemented tools, were evaluated in an iterative and informal way during the lifetime of the project. This included an exercise to assess the quality and completeness of archived web material. In this context research questions such as (i) 'What percentage of Belgian history is lost as a result of the lack of a Belgian web-archive?', (ii) 'What websites resisted time and are still online?' and (iii) 'How much of the Belgian web of the past can be reconstructed through other web-archives or using other 'web archaeology'-techniques?' were explored. Results were (amongst others) presented in the conference paper 'Unearthing the Belgian web of the 1990's: a digitised reconstruction' which was presented at 'The Web That Was: Archives, Traces and Reflections', the 3rd RESAW Conference (Amsterdam, 19-21 June 2019). For other evaluation activities the personas created in the Corpus project and outlined in the deliverable 'Le projet Corpus et ses publics potentiels. : Une étude prospective sur les besoins et les attentes des futurs usagers' (see https://hal-bnf.archives-ouvertes.fr/hal-01739730/document) were used. As such, these five personas, initially described in order to help with the identification of potential users were used to assess current access methods to, and analysis tools for, web archives.

In terms of recommendations for sustainable web archiving, a number of actions have been undertaken. First of all, several recommendations have been made to revise and to modify the legal deposit legislation

in Belgium. Secondly, an in-depth report about legal considerations concerning access to web archives has been drafted. Additionally, a list of operational Frequently Asked Questions (FAQ), which can be used by the State Archives and KBR as guidelines for personal data protection in the context of web archiving, were prepared. Thirdly, decision trees for assessing copyright for web archiving at both the selection and access stage have been created. Fourthly, a FAQ on personal data protection has been drafted to help KBR and AGR employees understand the challenges involved in web archiving. A business model was developed for KBR and the State Archives and operational procedures covering the entire web archiving workflow were also outlined.

Keywords

- web archiving
- digital humanities
- digital preservation
- online information
- Internet

PROMISE: Het bewaren van veelsoortige online informatie: naar een Belgische strategie
SAMENVATTING

Context

Het internet is geëvolueerd tot een communicatiemiddel dat centraal staat in ons leven van elke dag en ook voor erfgoed is het een bijzonder waardevol instrument geworden. Als collecties van data en documenten vertegenwoordigen websites belangrijk materiaal dat dient te worden gearchiveerd. Net als voor andere publicaties en archieven, waarvan de preservering wordt gewaarborgd door de wetgeving op het wettelijk depot en de archiefwet, moet er voor het Belgische web een beleid van bewaring op lange termijn worden ontwikkeld.

Doelstellingen

Het project PROMISE werd gelanceerd om een antwoord te formuleren op de dringende vraag hoe de preservering van het Belgische web - als een belangrijk onderdeel van de Belgische geschiedenis - voor toekomstige generaties moet worden aangepakt. In het PROMISE project werden vier doelstellingen voor op gesteld:

- Actuele beste praktijken op het vlak van webarchivering ontwikkelen
- Een Belgisch beleid voor webarchivering op federaal niveau uitstippelen
- Webarchivering, toegang tot en gebruik van de pilootversie van het Belgische webarchief voor wetenschappelijk onderzoek sturen
- Aanbevelingen formuleren voor een duurzame Belgischewebarchiveringstrategie

Methodologie

Op methodologisch vlak begon het project PROMISE data te verzamelen (uit wetteksten, bestaande initiatieven, de rol van instellingen) door middel van een literatuuroverzicht en semigestructureerde interviews met vertegenwoordigers van buitenlandse webarchivering initiatieven. Op basis van de bijeengebrachte informatie werd een realistische strategie en beleid uitgewerkt met het oog op het verzamelen en bewaren van het Belgische web. Dit omvatte een uitgebreide kostenberekening op basis van realistische scenario's voor webarchivering. Tevens werd een enquête gehouden om de vereisten van gebruikers van online archieven te analyseren; met de resultaten daarvan werd ook rekening gehouden bij het opstellen van een strategie.

Vervolgens werd het ontwikkelde model gevalideerd door een piloot-webarchief samen te stellen, bestaande uit de selectie en harvesting van de inhoud en het openstellen van de collecties voor toegang en gebruik. Tot slot formuleerde het project PROMISE een aantal aanbevelingen voor de implementatie van een duurzame dienst voor webarchivering in België, inclusief wettelijke en operationele perspectieven. Op wettelijk vlak werd een grondige studie gevoerd naar de wetgeving voor de bescherming van persoonsgegevens op Europees (AVG) en Belgisch niveau. Het operationele aspect

omvatte de ontwikkeling van een zakenmodel op basis van het Service-Dominant Business Model en het vastleggen van operationele procedures, waarbij rekening werd gehouden met de onderzoeksresultaten die tijdens PROMISE werden verkregen.

Conclusies

Het project PROMISE leverde een gedetailleerd rapport op met een stand van zaken inzake internationale "beste praktijken" voor webarchivering, rekening gehouden met wettelijke, operationele en technische aspecten. De voornaamste conclusies van deze projectfase werden ook gepubliceerd als onderzoeksartikel (Vlassenroot et al., 2019, zie https://link.springer.com/article/10.1007/s42803-019-00007-7).

Voor de definitie van een Belgisch, federaal webarchiveringsbeleid werd een rapport geschreven over het rechtskader van Belgische online informatie. Tevens werd een beleidsnota opgesteld voor de raad van bestuur van zowel het Rijksarchief als KBR. Deze nota omvatte de verschillende fasen in het proces van webarchivering (selectie, invoer, incorporatie, preservering, toegang …) evenals een gedetailleerde kostenanalyse voor een functioneel webarchief op basis van realistische scenario's. Er werd ook een onderzoek gevoerd naar de vereisten van de gebruikers van webarchieven; de resultaten daarvan werden gepubliceerd in de vorm van een interactief dashboard (zie https://public.tableau.com/profile/eveline.vlassenroot#!/vizhome/PReservingOnlineMultipleInformationtowardsaBelgianstrategy/PReservingOnlineMultipleINformationtowardsaBelgianstrategy).

De pilootfase van het project behelsde de selectie van te archiveren webinhoud en van de daartoe vereiste instrumenten, de harvesting van die inhoud en het beheer van de toegang tot die collecties. Er werden zogenaamde seed lists (i.e. lijsten met relevante URL's) opgesteld voor het Rijksarchief en KBR. De websites op deze seed lists werden gecapteerd en opgeslagen met de open source software Heritrix, wat resulteerde in een collectie van WARC-bestanden. Het webarchief werd toegankelijk gemaakt met behulp van de open source software PyWB waarvan een instance werd geïnstalleerd op de servers van het Rijksarchief en KBR. De gecapteerde inhoud en de gebruikte instrumenten waren gedurende het project het voorwerp van een iteratieve en informele evaluatie. Dit omvatte een oefening ter evaluatie van de kwaliteit en de volledigheid van het gearchiveerde webmateriaal. In deze context werden antwoorden gezocht op onderzoeksvragen als (i) 'Welk percentage van de Belgische geschiedenis is verloren gegaan als gevolg van het ontbreken van een Belgisch webarchief?', (ii) 'Welke websites bleken bestand tegen de tand des tijds en staan nog steeds online?' en (iii) 'Hoeveel van het oude Belgische web kan worden gereconstrueerd via andere webarchieven of via 'webarcheologie technieken'?'. De resultaten werden (onder andere) gepubliceerd in de conferentie paper 'Unearthing the Belgian web of the 1990's: a digitised reconstruction' die werd voorgesteld ter gelegenheid van de 3de RESAW-conferentie 'The Web That Was: Archives, Traces and Reflections' (Amsterdam, 19-21 juni 2019). Voor andere evaluatieactiviteiten werd gebruik gemaakt van de personae die werden gecreëerd in het Corpus-project en geschetst in de deliverable 'Le projet Corpus et ses publics potentiels : Une étude prospective

sur les besoins et les attentes des futurs usagers' (zie https://hal-bnf.archives-ouvertes.fr/hal-01739730/document). Deze vijf personae, die aanvankelijk werden beschreven om te helpen met de identificatie van potentiële gebruikers, werden gebruikt ter evaluatie van actuele methoden van toegang tot, en analyse-instrumenten voor, webarchieven.

Op het vlak van aanbevelingen voor duurzame webarchivering werd een aantal acties ondernomen. Om te beginnen werden verschillende aanbevelingen geformuleerd om de Belgische wetgeving inzake het wettelijk depot te herzien en te wijzigen. Vervolgens werd een grondig rapport opgemaakt met juridische beschouwingen betreffende de toegang tot webarchieven. Bovendien werd een lijst met operationele vaak gestelde vragen (Frequently Asked Questions, FAQ) opgesteld die het Rijksarchief en KBR kunnen gebruiken als richtsnoeren voor de bescherming van persoonsgegevens in de context van webarchivering. Ten derde werden "besluitvormingsbomen" gecreëerd voor de evaluatie van copyright voor webarchivering in zowel de selectie- als de toegangsfase. Ten vierde werd een lijst met FAQ inzake de bescherming van persoonsgegevens opgesteld om de werknemers van KBR en het Rijksarchief te helpen de uitdagingen in verband met webarchivering te begrijpen. Er werd een bedrijfsmodel opgesteld voor KBR en het Rijksarchief en er werden ook operationele procedures geschetst van de volledige werkstroom voor webarchivering.

Trefwoorden

- webarchivering
- digital humanities
- digitale duurzaamheid
- online informatie
- internet

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), se sont chargées d'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