PROMISE: Preserving Online Multiple Information: towards a Belgian strategy

SUMMA

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