

BelgicaWeb

Sustaining access to Belgium's born-digital heritage

Contract - B2/233/P2/BelgicaWeb

SUMMARY

Context, Objectives and Methodology

The web has become an essential environment in which contemporary social, cultural, political and scientific life is documented. Websites, online news platforms, blogs and social media constitute valuable sources for understanding recent history, yet they are highly fragile and can disappear rapidly. While significant progress has been made in preserving web content, major challenges remain regarding discovery, access, reuse, legal compliance and long-term sustainability.

The BelgicaWeb project (2024–2026), funded by BELSPO, brought together KBR, Ghent University and the University of Namur. Building on the foundations laid by the PROMISE and BESOCIAL projects, BelgicaWeb aimed to make Belgium's born-digital heritage more Findable, Accessible, Interoperable and Reusable (FAIR).

The project combined the expertise of several research centres. At Ghent University, mict led user requirements research, stakeholder engagement and participatory design activities, GhentCDH contributed expertise in digital humanities and research use cases, and IDLab focused on technical development and metadata enrichment. At the University of Namur, CRIDS provided expertise on copyright, data protection, artificial intelligence regulation and broader legal and ethical issues. KBR coordinated the project and led activities related to web archiving, collection development and access infrastructure.

BelgicaWeb pursued six objectives: developing sustainable approaches to access born-digital heritage; expanding Belgian web and social media collections; improving discoverability through metadata enrichment; designing an access platform and API; examining legal and ethical frameworks; and promoting awareness and use of born-digital heritage collections.

The project adopted an interdisciplinary and user-centred methodology combining digital preservation, digital humanities, computer science, archival science and legal research. User requirements were gathered through surveys, workshops and consultations involving researchers, heritage professionals and web archiving specialists. Two advisory structures supported the project. The Follow-up Committee provided periodic strategic advice and reviewed project progress, while the Expert Panel played a more active co-creation role by contributing to requirements gathering, seed list development, collection evaluation, quality assurance and access platform design.

Results, Main Conclusions and Recommendations

BelgicaWeb demonstrated that researchers increasingly view web archives as research infrastructures rather than preservation repositories alone. Users highlighted the need for transparent collection development, rich metadata, advanced search capabilities, downloadable datasets and programmatic access through APIs, while also stressing the importance of supporting both exploratory and computational research.

The project developed methodologies for building born-digital heritage collections, including seed list development, harvesting workflows and quality assurance procedures. The Expert Panel contributed significantly to improving the diversity and relevance of the collections. Browsertrix was selected as the primary harvesting solution because of its open-source nature and support for modern web technologies. The project also confirmed the continuing challenges associated with social media archiving, particularly due to technical, legal and platform-related constraints.

A major result was the development of a prototype access platform supporting collection discovery, search and browsing for different user groups. BelgicaWeb also designed and implemented an API to facilitate machine-readable access to archived content and metadata. The technical infrastructure was developed using open-source technologies, including Symfony and SolrWayback.

The project further developed enrichment workflows generating additional metadata such as quality indicators, geolocation information and copyright-related markers. It also explored and prototyped artificial-intelligence-based enrichment approaches including language identification, thematic categorisation, named entity recognition, automated summarisation and image description. These techniques were found to improve discovery and reuse when applied transparently and in combination with human expertise.

Legal research highlighted the need to balance openness and research reuse with compliance with copyright and data protection legislation. Particular attention was devoted to copyright, legal deposit, privacy, text and data mining, licensing approaches and emerging artificial intelligence regulation. A key finding was that legal and ethical considerations should be integrated throughout the archival lifecycle rather than addressed only after collections have been created.

BelgicaWeb demonstrated that sustainable access depends not only on technology but also on governance, institutional commitment, interdisciplinary cooperation and continued engagement with user communities. The project showed that web archives are increasingly evolving into research infrastructures supporting discovery, analysis and reuse. Its outputs, including user requirements studies, collection development methodologies, experimental collections, a prototype access platform, an API, enrichment workflows and legal recommendations, provide a foundation for future access to Belgium’s web and social media heritage.

Main Recommendations

- Continue developing web archives as sustainable research infrastructures.
- Maintain active involvement of user communities through co-creation and iterative evaluation.
- Invest in metadata enrichment and computational access services, including APIs.
- Integrate legal and ethical considerations throughout the archival lifecycle.
- Strengthen institutional collaboration and long-term governance arrangements.
- Continue exploring responsible applications of artificial intelligence for archive discovery and reuse.

Keywords: Web Archiving; Born-Digital Heritage; Research Infrastructure; FAIR Data; Digital Preservation