

BelgicaWeb

Assurer un accès durable au patrimoine nativement numérique de la Belgique

RESUME

Contexte, objectifs et méthodologie

Le web constitue aujourd’hui un espace essentiel où se documentent les activités sociales, culturelles, politiques et scientifiques contemporaines. Les sites web, plateformes d’information en ligne, blogs et réseaux sociaux représentent des sources précieuses pour comprendre l’histoire récente, mais ils demeurent particulièrement fragiles et susceptibles de disparaître rapidement. Malgré les avancées importantes réalisées dans la préservation du web, des défis majeurs subsistent en matière de découverte, d’accès, de réutilisation, de conformité juridique et de durabilité à long terme.

Le projet BelgicaWeb (2024–2026), financé par BELSPO, a réuni la KBR, l’Université de Gand et l’Université de Namur. Dans la continuité des projets PROMISE et BESOCIAL, BelgicaWeb visait à rendre le patrimoine nativement numérique belge davantage conforme aux principes FAIR (Facile à trouver, Accessible, Interopérable et Réutilisable).

Le projet a mobilisé l’expertise de plusieurs centres de recherche. À l’Université de Gand, mict a dirigé les travaux relatifs aux besoins des utilisateurs, à l’implication des parties prenantes et à la conception participative. GhentCDH a apporté son expertise en humanités numériques et en usages de recherche, tandis qu’IDLab s’est concentré sur le développement technique et l’enrichissement des métadonnées. À l’Université de Namur, le CRIDS a fourni une expertise en matière de droit d’auteur, de protection des données, de réglementation de l’intelligence artificielle et de questions juridiques et éthiques. La KBR a assuré la coordination générale du projet ainsi que les activités liées à l’archivage du web, au développement des collections et aux infrastructures d’accès.

BelgicaWeb poursuivait six objectifs : développer des approches durables pour l’accès au patrimoine nativement numérique ; enrichir les collections belges du web et des médias sociaux ; améliorer leur découvrabilité grâce à l’enrichissement des métadonnées ; concevoir une plateforme d’accès et une API ; analyser les cadres juridiques et éthiques applicables ; et promouvoir la connaissance et l’utilisation des collections de patrimoine nativement numérique.

Le projet a adopté une méthodologie interdisciplinaire et centrée sur les utilisateurs, combinant préservation numérique, humanités numériques, informatique, archivistique et recherche juridique. Les besoins des utilisateurs ont été identifiés au moyen d’enquêtes, d’ateliers et de consultations associant chercheurs, professionnels du patrimoine et spécialistes de l’archivage du web. Deux structures consultatives ont accompagné le projet. Le Comité de suivi a fourni des avis stratégiques périodiques et évalué l’avancement du projet, tandis que le Panel d’experts a joué un rôle plus actif de co-crédation en contribuant à l’identification des besoins, au développement des collections, à leur évaluation, au contrôle qualité et à la conception de la plateforme d’accès.

Résultats, principales conclusions et recommandations

BelgicaWeb a montré que les chercheurs considèrent de plus en plus les archives du web comme des infrastructures de recherche plutôt que comme de simples dépôts de préservation. Les utilisateurs ont souligné l'importance d'une politique transparente de développement des collections, de métadonnées riches, de fonctionnalités de recherche avancées, de jeux de données téléchargeables et d'un accès programmatique via des API, tout en insistant sur la nécessité de soutenir à la fois la recherche exploratoire et les approches computationnelles.

Le projet a développé des méthodologies pour la constitution de collections de patrimoine nativement numérique, notamment des listes de départ, des workflows de collecte et des procédures de contrôle de qualité. Le Panel d'experts a contribué de manière significative à améliorer la diversité et la pertinence des collections. Browsertrix a été retenu comme principal outil de collecte en raison de son caractère open source et de sa capacité à prendre en charge les technologies web modernes. Le projet a également confirmé les défis persistants liés à l'archivage des médias sociaux, notamment en raison de contraintes techniques, juridiques et liées aux plateformes.

L'un des principaux résultats du projet est le développement d'un prototype de plateforme d'accès permettant la découverte, la recherche et la consultation des collections pour différents groupes d'utilisateurs. BelgicaWeb a également conçu et mis en œuvre une API facilitant l'accès automatisé aux contenus archivés et à leurs métadonnées. L'infrastructure technique repose sur des technologies ouvertes telles que Symfony et SolrWayback.

Le projet a par ailleurs développé des workflows d'enrichissement produisant des métadonnées complémentaires telles que des indicateurs de qualité, des informations géographiques ou des marqueurs liés au droit d'auteur. Il a également exploré et prototypé plusieurs approches d'enrichissement fondées sur l'intelligence artificielle, notamment l'identification des langues, la catégorisation thématique, la reconnaissance d'entités nommées, le résumé automatique et la description d'images. Ces approches peuvent améliorer la découverte et la réutilisation des archives web lorsqu'elles sont appliquées de manière transparente et en complément de l'expertise humaine.

Les recherches juridiques ont mis en évidence la nécessité de concilier ouverture, réutilisation scientifique et respect de la législation relative au droit d'auteur et à la protection des données. Une attention particulière a été accordée au dépôt légal, au droit d'auteur, à la protection de la vie privée, au text and data mining, aux modèles de licences et aux réglementations émergentes relatives à l'intelligence artificielle. Une conclusion importante du projet est que les considérations juridiques et éthiques doivent être intégrées à l'ensemble du cycle de vie archivistique.

BelgicaWeb a démontré que l'accès durable dépend non seulement de la technologie, mais également de la gouvernance, de l'engagement institutionnel, de la coopération interdisciplinaire et du dialogue continu avec les communautés d'utilisateurs. Le projet a montré que les archives du web évoluent progressivement vers de véritables infrastructures de recherche favorisant la découverte, l'analyse et la réutilisation des données. Ses résultats, notamment les études de besoins des utilisateurs, les méthodologies de développement des collections, les collections expérimentales, le prototype de plateforme d'accès, l'API, les workflows d'enrichissement et les recommandations juridiques, constituent une base solide pour l'avenir de l'accès au patrimoine web et aux archives des médias sociaux en Belgique.

Principales recommandations

- Poursuivre le développement des archives web comme infrastructures de recherche durables.
- Maintenir l'implication active des communautés d'utilisateurs par la co-crédation et l'évaluation continue.
- Investir dans l'enrichissement des métadonnées et les services d'accès computationnel, notamment les API.
- Intégrer les considérations juridiques et éthiques tout au long du cycle de vie archivistique.
- Renforcer la collaboration institutionnelle et les mécanismes de gouvernance à long terme.
- Continuer à explorer les usages responsables de l'intelligence artificielle pour la découverte et la réutilisation des archives.

Mots-clés: Archivage du web ; Patrimoine nativement numérique ; Infrastructure de recherche ; Données FAIR ; Préservation numérique