The first steps towards a Belgian web archive: a federal strategy

WAC – Zagreb
June 6th 2019

Friedel Geeraert  Royal Library (KBR) and State Archives of Belgium
Sébastien Soyez  State Archives of Belgium
Overview

The PROMISE project
The strategy
Lessons learnt
Next steps
I The PROMISE project

AIM

To develop a federal strategy for the preservation of the Belgian web

PARTNERS

GHENT UNIVERSITY

UNIVERSITÉ DE NAMUR

HE2B

belspo

TIMING

July 2017 – December 2019
The PROMISE project

- Literature review
- Interviews with representatives of web archiving initiatives

Identify best practices in the field of web archiving
II The strategy
Selection

- .be
- .vlaanderen
- .brussels
- .gent

**.org**
**.com**
...

**ccTLDs**
**.fr**
**.nl**
...

**gTLDs**

IF

Web content created, produced, published, ... on the Belgian territory

Websites are registered by Belgians

Content concerns Belgium or its general affairs
Selection

**BROAD CRAWL**
- Sampling of the Belgian web
- Superficial capture
- Collected once a year
- Problem: no access to full list of Belgian domain names

**SELECTIVE CRAWLS**
- Complete capture of web content
- More frequent captures
- Federal institutions
- Ministerial cabinets, ministers/secretaries of state
- Public organisms with link to federal level
- Provinces, the regions and the communities
- Projects funded by BELSPO

601 websites

- Special heritage collections
- Contemporary collections
- Spanish, Italian and Portuguese communities

928 websites
1416 web pages
37 sections of websites
Collecting

Server for collecting

XML/OCLC Descriptive Metadata

WARC Web Archive + Technical Metadata

HERITRIX Crawler robots

KBR

KBR

KBR
Quality control

- Lack of automated tools
- Prototype visual correspondence
Access

Web collection AGR/KBR

KBR & AGR (XML/OCLC)

KBR (XML/MARC21)

Belgian Web Archives

AGR (XML/EAD)

Users
Access

Law on Archives

Legal Deposit Law

Copyright legislation

Privacy Legislation (GDPR)

Illegal content
III Lessons learnt

Selection: from book to web

A lot of web pages and Wikipedia

What is useful content?

DNS Belgium

Preservation
III Lessons learnt

- Web crawling takes time
- Setting crawl parameters = trial and error
- Quality control = pain point
- Estimations: time and cost
- Too ambitious?
IV Next steps

- Validation of shared strategy
- Access to archived web content
- Recommendations and procedures
- Web archiving = part of KBR strategy 2019-2021
- October 18th - Colloquium ‘Saving the web: the promise of a Belgian web archive’
Thank you!

Friedel Geeraert
Researcher on PROMISE project
friedel.geeraert@kbr.be