The promise of web archiving in Belgium

Emmanuel Di Pretoro - URF-SID, Haute-École Bruxelles-Brabant
Friedel Geeraert - Royal Library and State Archives
Alejandra Michel - NADI/CRIDS, UNamur
Eveline Vlassenroot - imec-mict-UGent

Overview

1. Introducing the PROMISE project
2. State of the art
    a. Selection
    b. Access
    c. Legal aspects
    d. Technical aspects
3. Survey on user requirements
4. Next steps
5. Q&A

1. Introducing the PROMISE project

(Preserving Online Multiple Information: Towards a Belgian Strategy)

What is web archiving?

Definition of web archiving by the International Internet Preservation Consortium

*"Web archiving is the process of **collecting** portions of the World Wide Web, **preserving** the collections in an archival format, and then serving the archives for **access and use**."*

(source: IIPC, 2018, *Web archiving*, retrieved from http://netpreserve.org/web-archiving/, last accessed on 16/05/2018)

Why archive the web?

The web is ever-present in daily life and communication

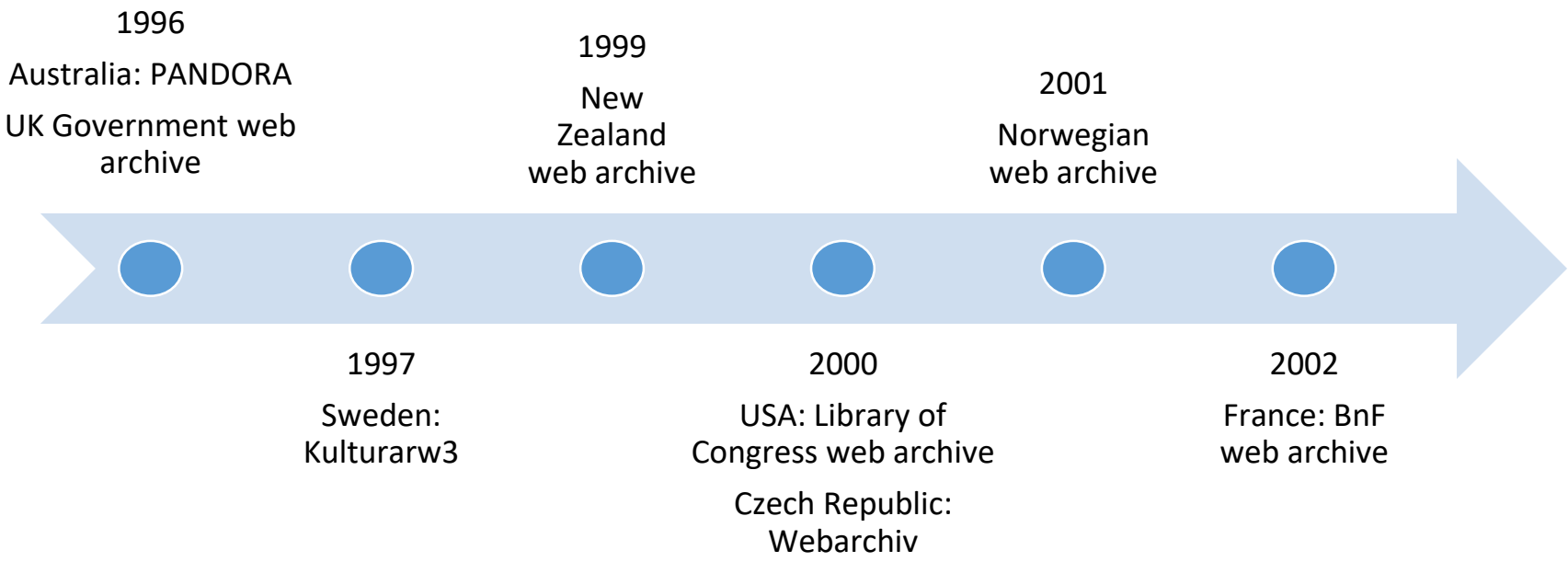
It increasingly holds the traces of our history

=> Goldmine of information?

Ephemeral nature of online information

=> Digital dark age?

International web archiving initiatives



Introducing the PROMISE project

belspo

- Challenge: the Belgian web is currently not systematically archived
- Collaborative strategy = innovation + long-term financial and operational benefits
- Project partners:

Archives de l'État
Rijksarchief
www.arch.be

Koninklijke Bibliotheek van België
Bibliothèque royale de Belgique

GHENT UNIVERSITY

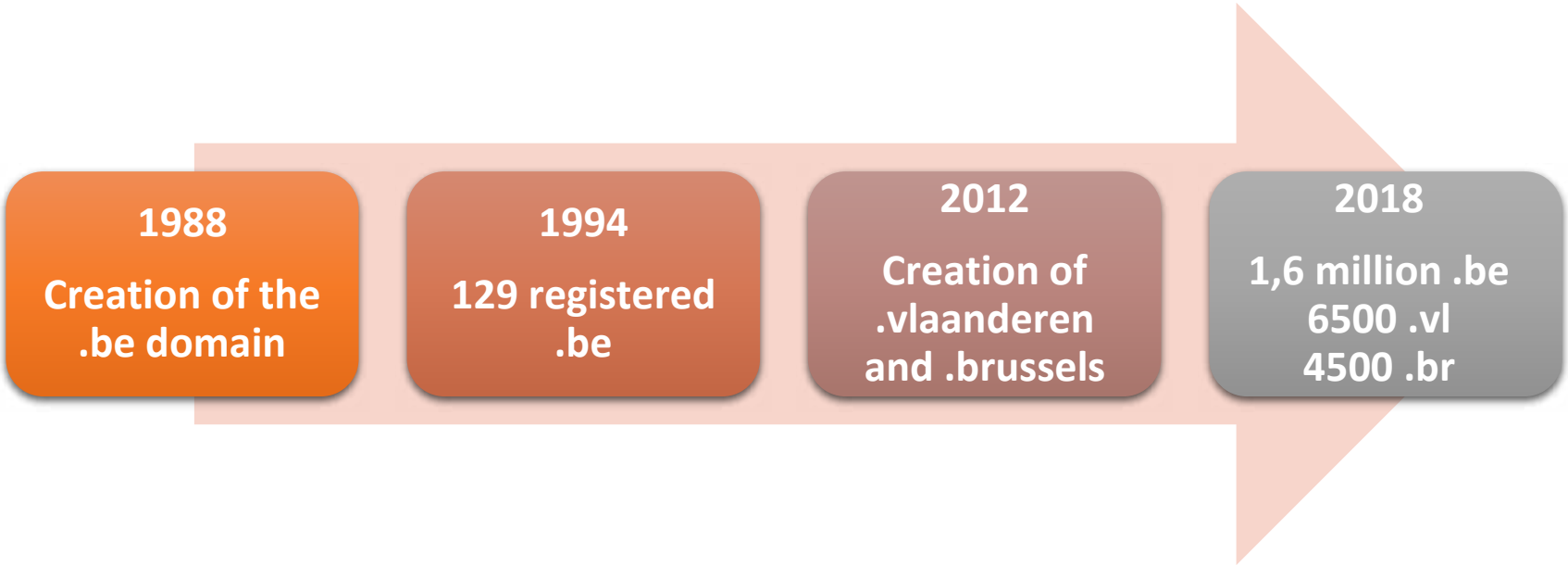
HE2B
HAUTE ÉCOLE
BRUXELLES-BRABANT

UNIVERSITÉ DE NAMUR

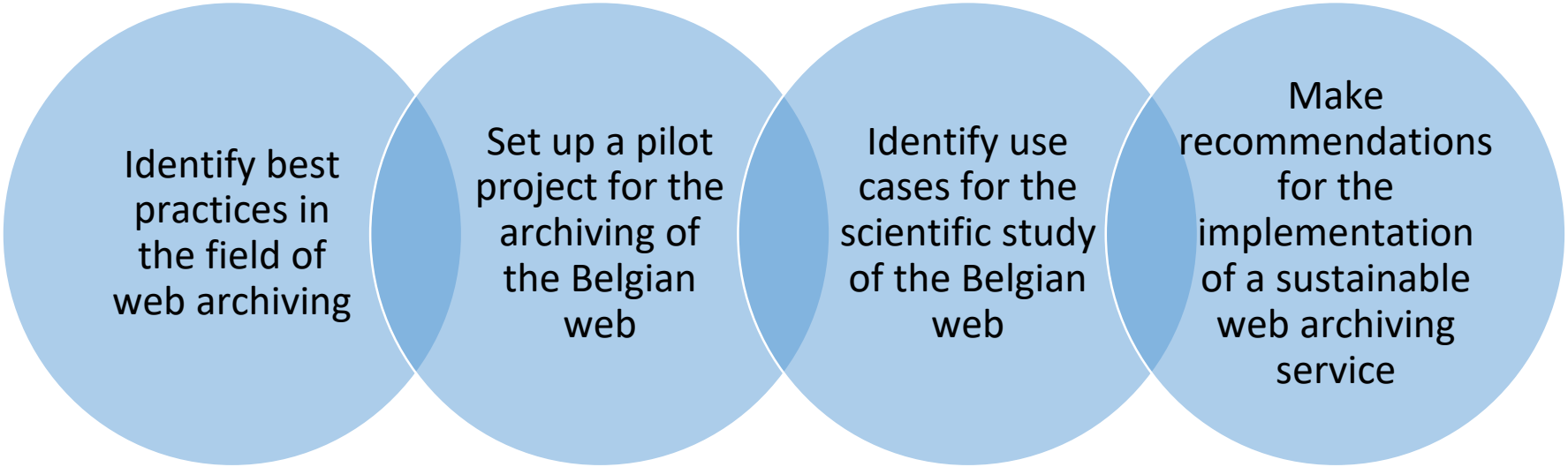
crids

- When: 2017 - 2019

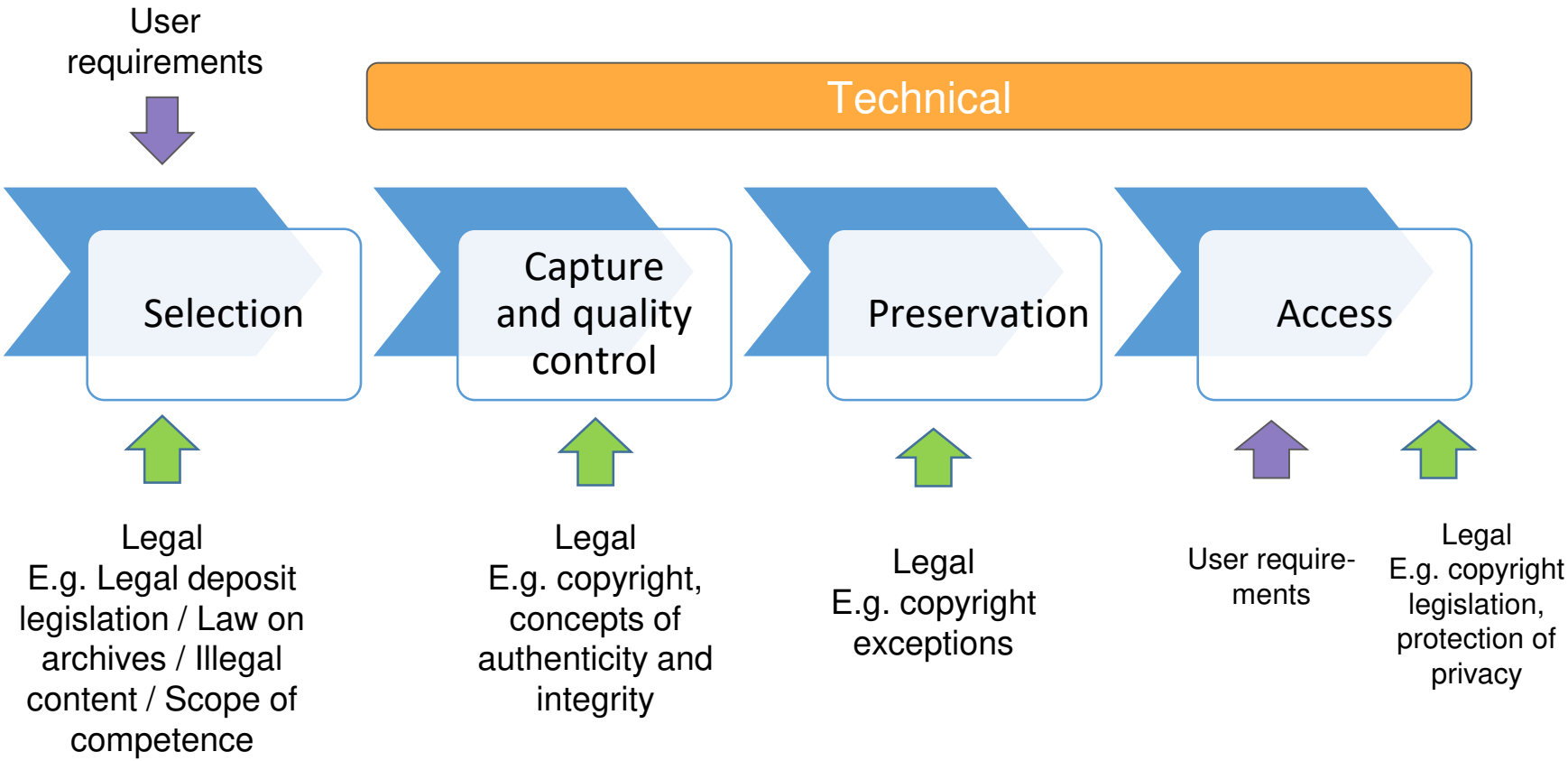
The Belgian web: a history



PROMISE project phases



How?


User requirements
Technical
Selection
Capture and quality control
Preservation
Access
Legal
E.g. Legal deposit legislation / Law on archives / Illegal content / Scope of competence
Legal
E.g. copyright, concepts of authenticity and integrity
Legal
E.g. copyright exceptions
User require-ments
Legal
E.g. copyright legislation, protection of privacy

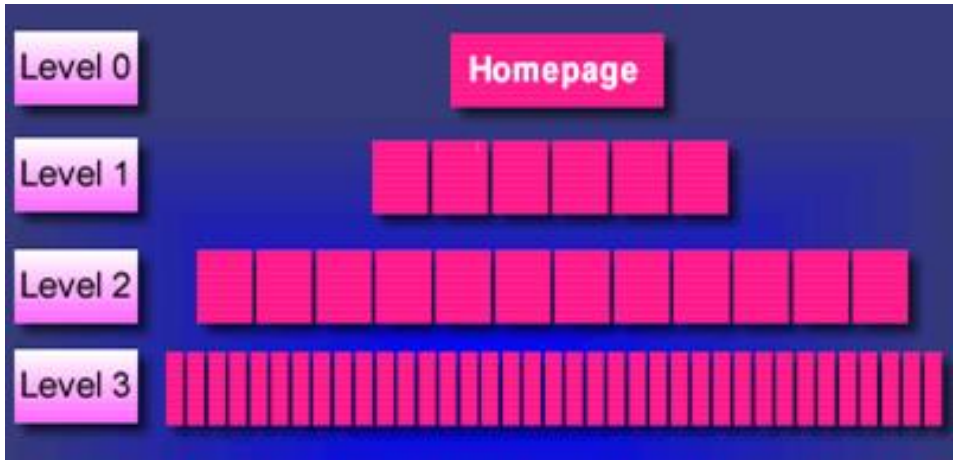
2. State of the art

KB
na
nationaal archief
netarkivet.dk
Library and Archives Canada
Bibliothèque et Archives Canada
Canada
The National Archives
BRITISH LIBRARY
BAnQ
Bibliothèque nationale de Luxembourg
Nationalbibliothéik
BnF
ina
National Library of Ireland
ARQUIVO.PT
Schweizerische Eidgenossenschaft
Confédération suisse
Confederazione Svizzera
Confederaziun svizra
Eidgenössisches Departement des Innern EDI
Schweizerische Nationalbibliothek NB

2.1 State of the art: selection

Selection policy

- Selection = URI + depth + frequency
- Legal framework
  - Libraries: link with legal deposit legislation
  - National Archives: link with ‘Law on legislation’



Selection policy: national libraries

BROAD CRAWLS
(superficial capture)

1. The national domain (top-level domain crawls)
2. Other websites that are considered interesting

SELECTIVE CRAWLS
(complete capture)

1. Themes
2. Events
3. Emergencies

Country	Institution	Broad crawls	Selective: thematic	Selective: events	Selective: other
Netherlands	Nat. Library	No	Yes	No	No
France	Nat. Library	No (representative sample)	Yes	Yes	Yes (emergencies)
United Kingdom	British Library	Yes (non-print legal deposit web archive)	Yes (open UK web archive)	Yes (open UK web archive)	Yes (emergencies- open UK web archive)
Luxembourg	Nat. Library	Yes	No	Yes	No
Denmark	Royal Danish Library	Yes	Yes	Yes	Yes (emergencies, research projects, videos)
Portugal	Arquivo.pt	Yes	No	Yes	No
Ireland	Nat. Library	Yes	Yes	Yes	No
Canada	Libr. & Arch. Canada	No (in preparation)	Yes	Yes	Yes (emergencies, risk of disappearing)
Canada	Nat. Libr. & Arch. Quebec	No	Yes	Yes	No
Switzerland	Nat. Libr.	No	Yes	Yes	No

Selection policy: national libraries

- Selective collections
  - General selection criteria
  - Themes (e.g. departments, post-truth, ...)
  - Events (e.g. Olympics, elections, …)
  - Emergencies (e.g. natural disasters, …)
- Multilingualism: different approaches

Selection policy: national archives

WEB ARCHIVING BY ARCHIVE PRODUCER

- Nationaal Archief NL: government institutions archive their web content and then transfer it to the Nationaal Archief

WEB ARCHIVING BY NATIONAL ARCHIVES

- Nationaal Archief NL: only for purchased archives
- UK National Archives
- Bibliothèques et Archives nat. Québec

Selection policy: social media

	Facebook	Twitter	Youtube	Instagram	Flickr
France (Nat. Libr.)	Not anymore	Yes	No	No	No
Denmark (Roy. Libr.)	Yes	Yes	Yes	Yes	No
Luxembourg (Nat. Libr.)	Yes	Yes	Yes	Yes	No
UK (British Library)	Yes	Yes	No	No	No
Ireland (Nat. Libr.)	No	Yes	Yes	No	Starting 2018
UK (Nat. Arch.)	No	Yes	Yes	No	No
Library and Archives (Canada)	Yes	Yes	Yes	Yes	No

2.2 State of the art: access

GHENT UNIVERSITY

Archives de l'État
Rijksarchief
www.arch.be

Overview of access methods

Country	Institution	Access method		Who has access?
		Open & freely accessible online	Physical access on location	
The Netherlands	National Library	No	Yes	Everyone with a paid library card. Big data researchers can gain access after a meeting and having signed a contract.
The Netherlands	National Archive	Yes (for websites with an 'open' status)	Yes (for websites with a 'restricted' or 'offline' status)	'Open' & 'offline' status websites: everybody. Some items are 'restricted', which means you need a special permission (a research proposal is required to obtain this permission or proof that the subject of the archived content is dead). Together with the special permission a signed form is needed stating you understand your own responsibilities under the privacy-law.
France	National Library	No	Yes (but also from within the 26 partner libraries)	Authorized users of the BnF (18 years or older and for university studies, professional or personal research. For the latter two categories, interviews are conducted before accreditation is given.)
Luxembourg	National Library	No	No	No public system yet.
UK	British Library	Yes (for the UK web archive)	Yes (for the legal deposit UK web archive and JISC domain dataset)	Everyone with a reader's pass.
UK	National Archives	Yes	No	Everyone
Denmark	Royal Danish Library	Yes (only for researchers conducting research on a Ph.D-level or above)	Yes (only for researchers)	Only for research purposes after filling an application form that needs to be evaluated.
Portugal	Foundation for Science and Technology	Yes	No	Everyone
Ireland	National Library	Yes	No	Everyone

When access is obtained..

2 challenges:

- Lack of descriptive metadata guidelines
- Lack of a clear understanding of user needs and behaviour

GHENT UNIVERSITY

Archives de l'État
Rijksarchief
www.arch.be

Search options

Country	Institution	Search options			
		URL	Full-text	Topical browsing	Alphabetic browsing
The Netherlands	National Library	Yes	No	No	No
The Netherlands	National Archive	No	No	No	No
France	National Library	Yes	Yes	Yes	No
Luxembourg	National Library	closed for public	closed for public	closed for public	closed for public
UK	British Library	Yes	Yes	Yes	No
UK	National Archives	Yes	Yes	No	Yes
Denmark	Royal Danish Library	Yes	Yes	No	No
Portugal	Foundation for Science and Technology	Yes	Yes	No	No
Ireland	National Library	Yes	Yes	No	Yes

Example: UK Government web archive

2.3 State of the art: legal aspects

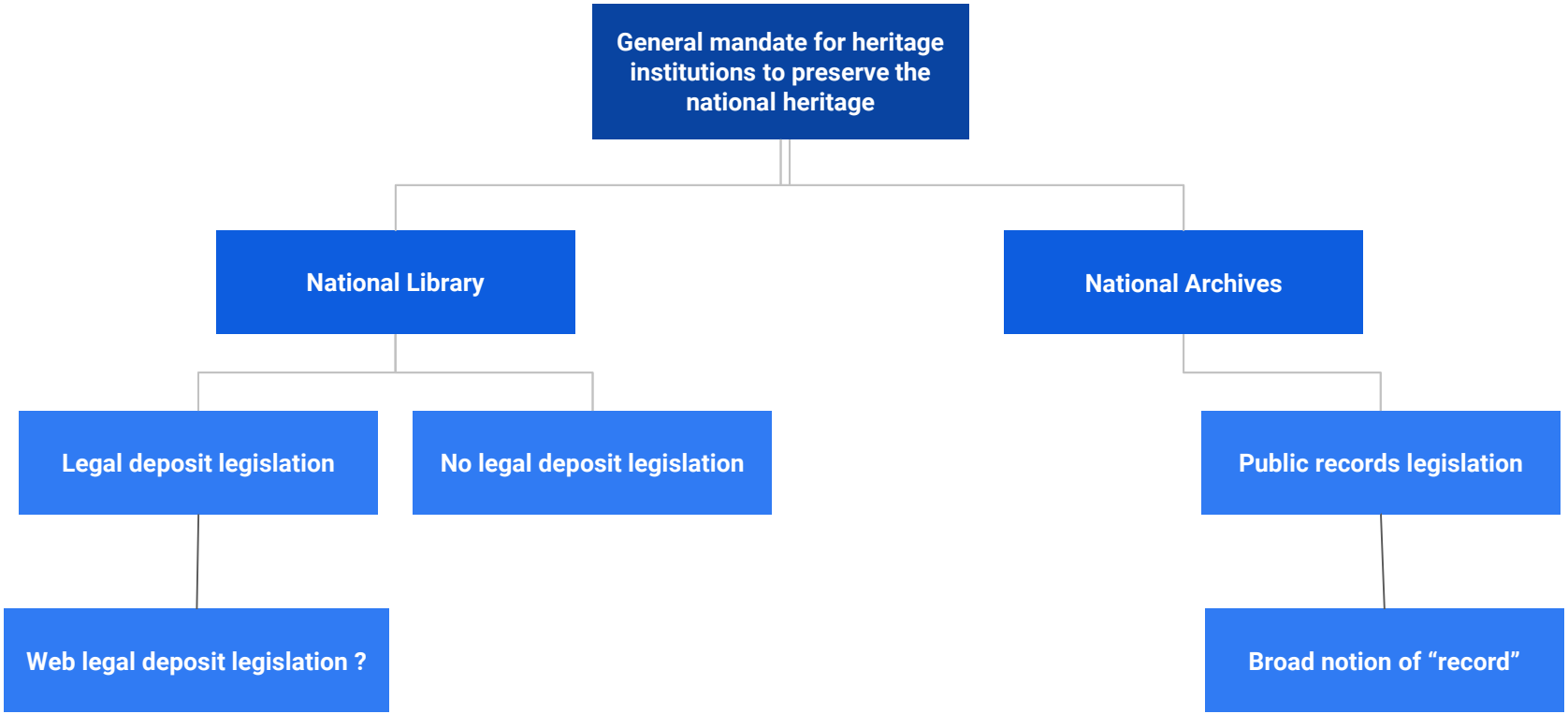
Legal aspects of web archiving

➔ **Web archives** are of the <u>utmost importance</u> to guarantee **the right of the public to information** (seek & receive information).

- (ECHR, *Times Newspapers Ltd c. United Kingdom (n° 2)*, 10 March 2009, §27).
- Thus **web archives** are **protected by Article 10 European Convention on Human Rights (ECHR)**.

<u>BUT</u> …. a lot of legal issues :

Which legal basis for web archiving ?



What are the advantages of a clear legal framework for web archiving ?

Heritage institutions

- **Legal certainty**
- **Simplification of web archiving activities**
  - legal obligation
  - copyright exception (prior authorization of copyright owner not required)
  - obligation to give the necessary passwords & access keys
- **Collaboration with domain names management bodies**
  - identification of websites owners

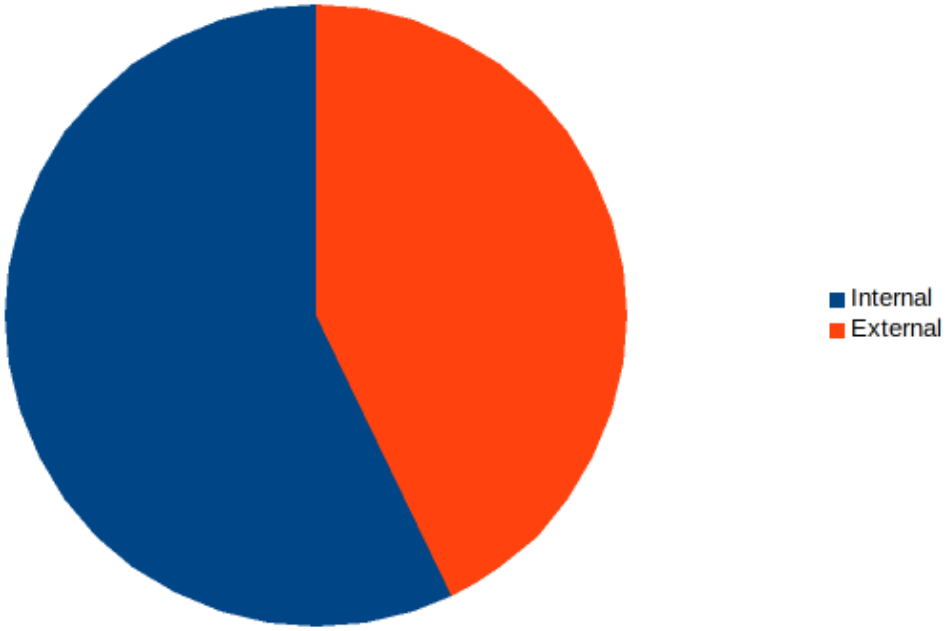
Websites owners or Copyright owners

- **Information on web archiving or web harvesting procedures**
- **Access embargo possibilities to protect their interests**

2.4 State of the art: technical aspects

Differents technical strategies to preserve the web

1. « Internal » solution
2. « External » solution



Architecture of a technical solution

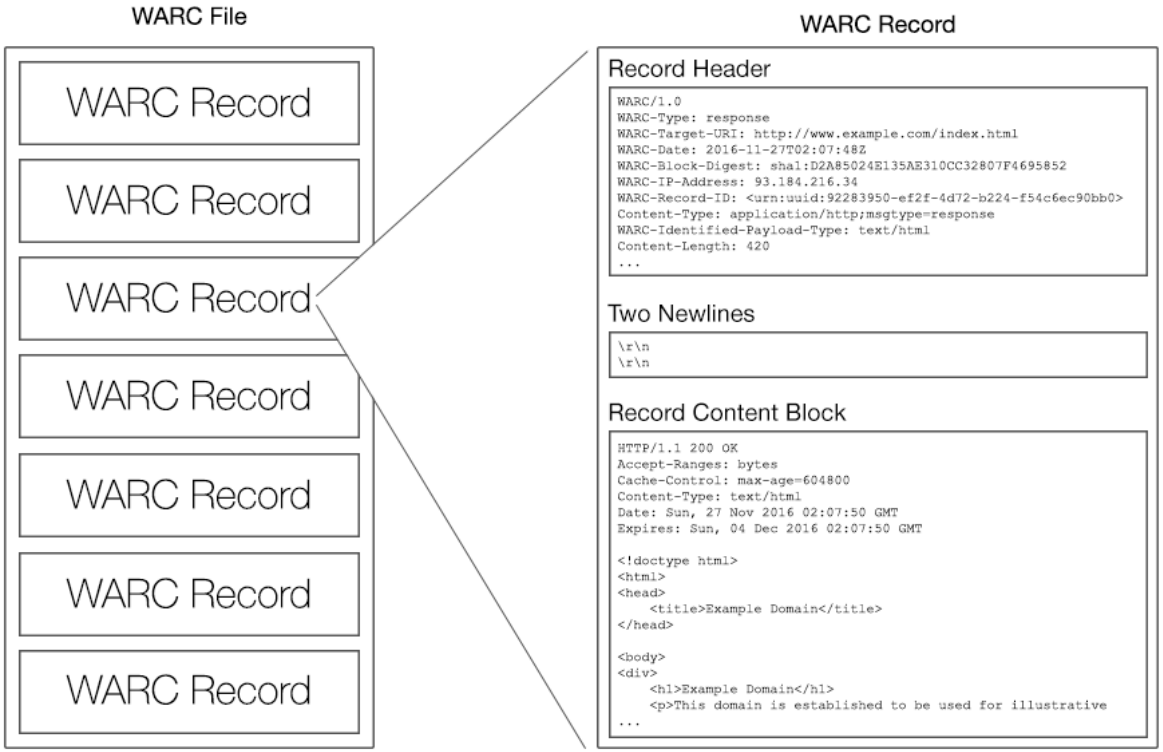
The solution should be able to answer at different questions:

1. How to collect a webpage or a website and store the results?
2. How to preserve a web archive in the long term?
3. How to give access to web archives?
4. How to facilitate the processing of web archives for the users?

List of technical solutions

- **Archive-It**
- **MirrorWeb**
- *Heritrix*
- *Wget*
- *OpenWayback*
- *Brozzler*
- *WebRecorder Player*
- ...

Brief presentation of WARC



[Source]

A Principle: *Recording all the HTTP interactions between the server (the website) and the client (the crawler, i.e. Heritrix)*

A Norm: ISO 28500

3. Survey user requirements

Survey on web archiving requirements

GHENT UNIVERSITY

promise

Survey on web archiving requirements

What are the requirements and needs of (potential) users of web archives? How do they, or do they want to, access, use and consult web archives?

This survey is part of the PROMISE project which is looking at how current best practices in web-archiving can be applied to the Belgian context.

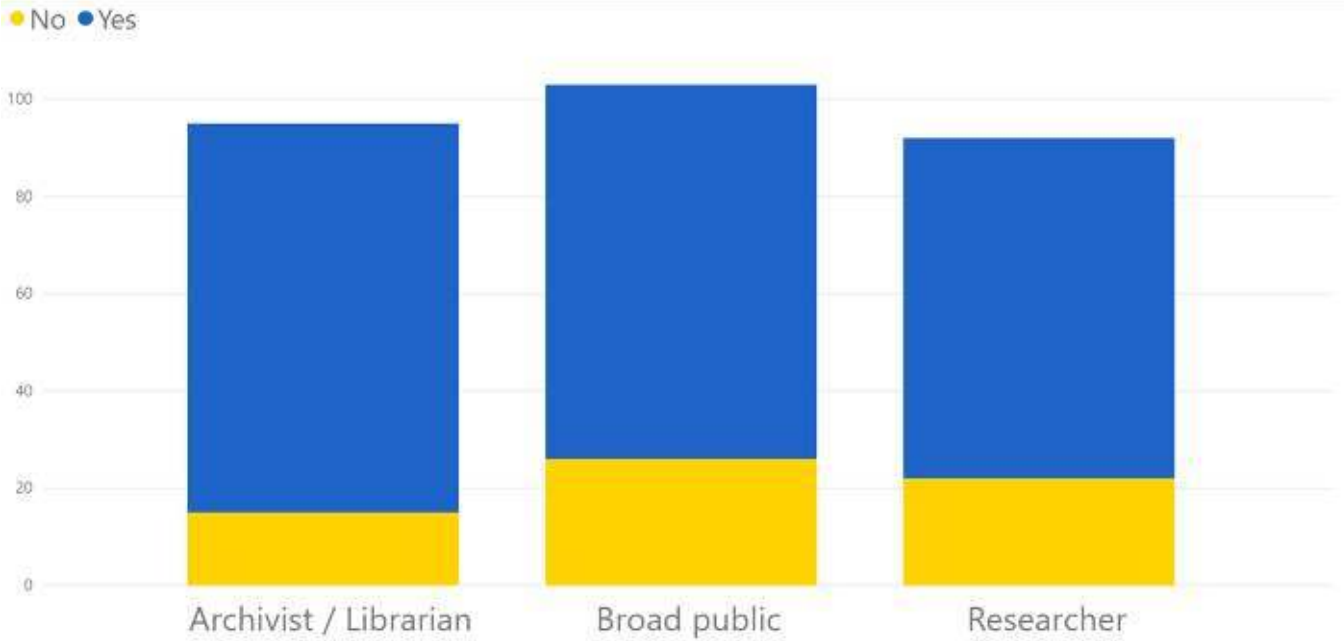
The PROMISE project is financed by the Belgian Science Policy Office (Belspo) as part of the BRAIN.be programme and is coordinated by the Royal Library. The State Archives, Ghent University (Research Group for Media, Innovation and Communication Technologies and Ghent Centre for Digital Humanities) and Univeristy of Namur (Research Centre in Information, Law and Society) and the university college Bruxelles-Brabant (Unité de Recherche et de Formation en Sciences de l'Information et de la Documentation) are partners in the project.

- Analysis of requirements related to selection and use of web archives
- Ran April 2018 - 31 May 2018
- Target 200+ respondents
- Local, regional, national and international focus

3 Target groups

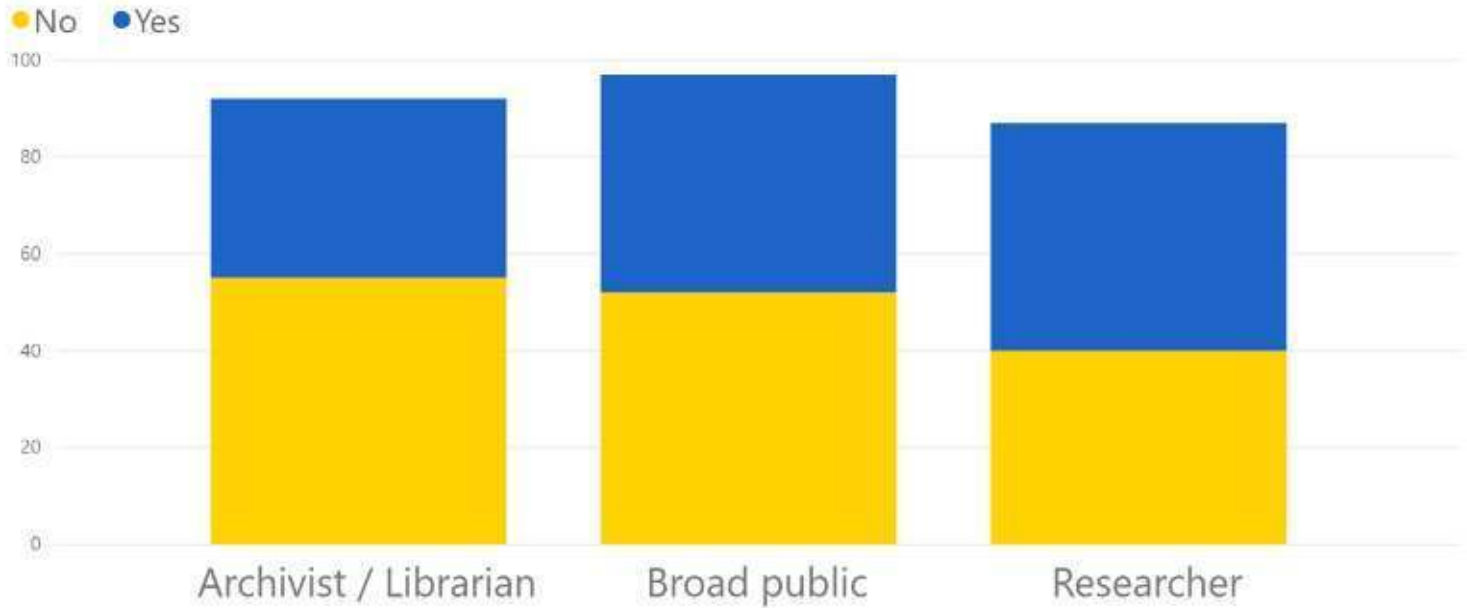
1. Research: students, academics, anyone involved in research more broadly
2. Archives, libraries, governmental institutions
3. General public

Are you aware of the existence of web archives?



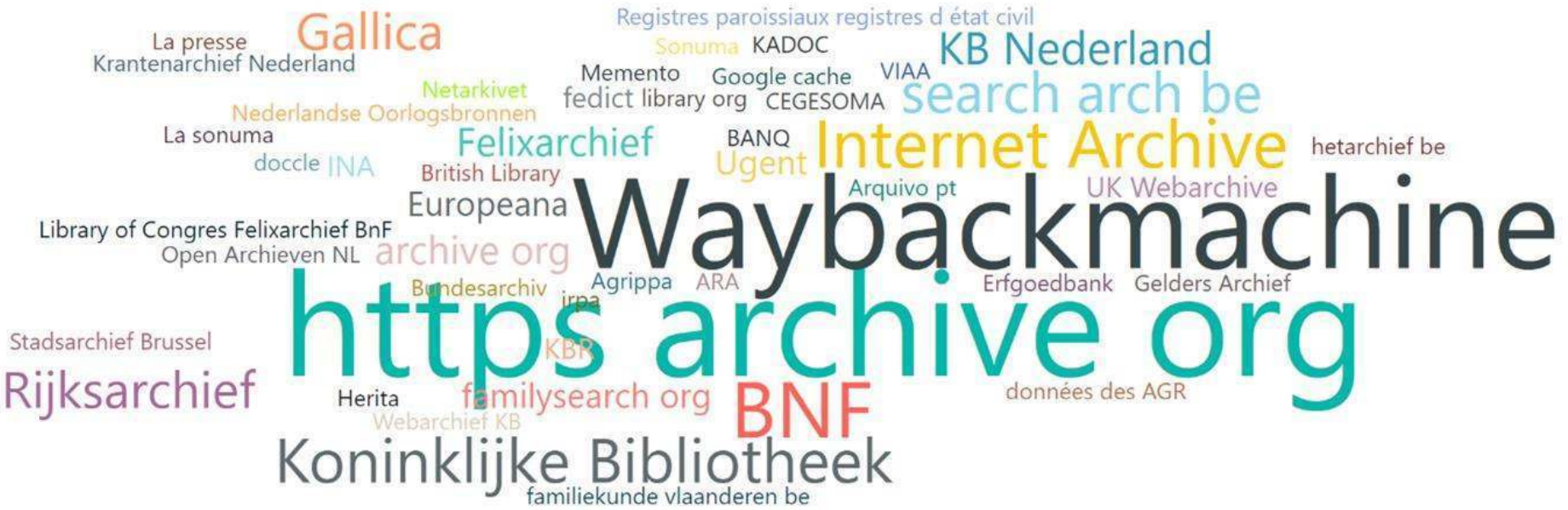
source: Preliminary results from survey launched in April 2018 with regards to the project PROMISE, N=276

Have you ever used a web archive?



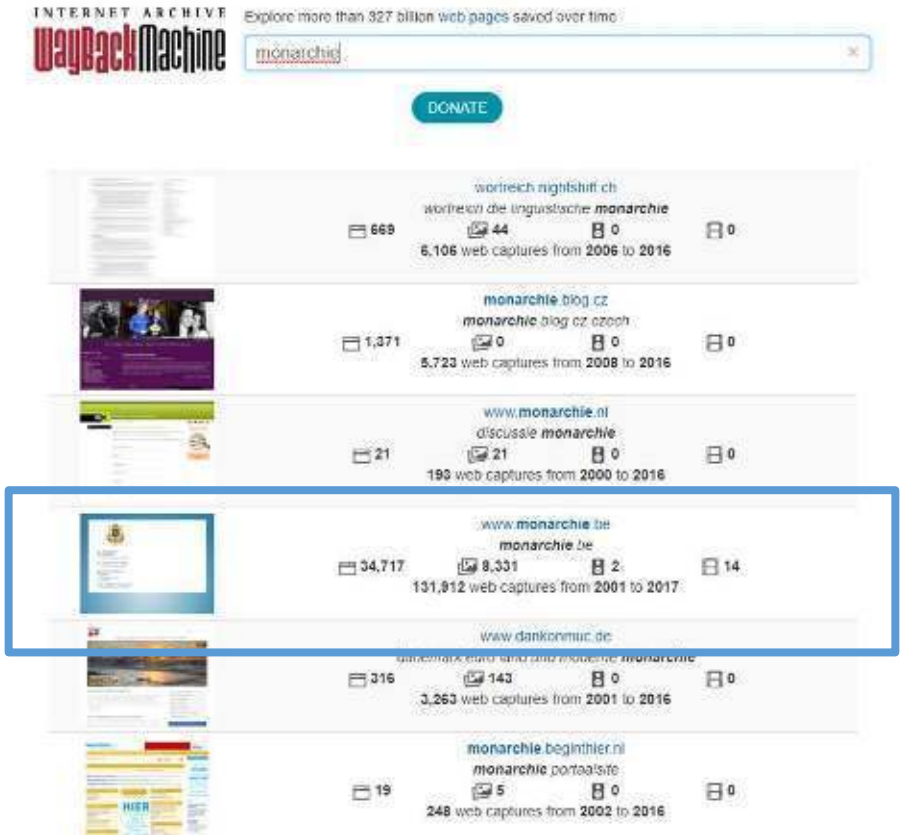
source: Preliminary results from survey launched in April 2018 with regards to the project PROMISE, N=276

Can you provide examples of web archives?



source: Preliminary results from survey launched in April 2018 with regards to the project PROMISE, N=276

Internet Archive - Wayback Machine



GHENT UNIVERSITY

Archives de l'État
Rijksarchief
www.arch.be

INTERNET ARCHIVE
WayBackMachine

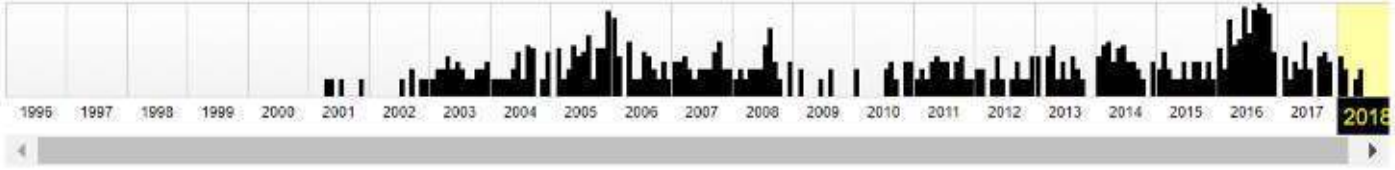
Explore more than 327 billion web pages saved over time

http://monarchie.be/

DONATE

Saved **721 times** between April 17, 2001 and May 13, 2018.

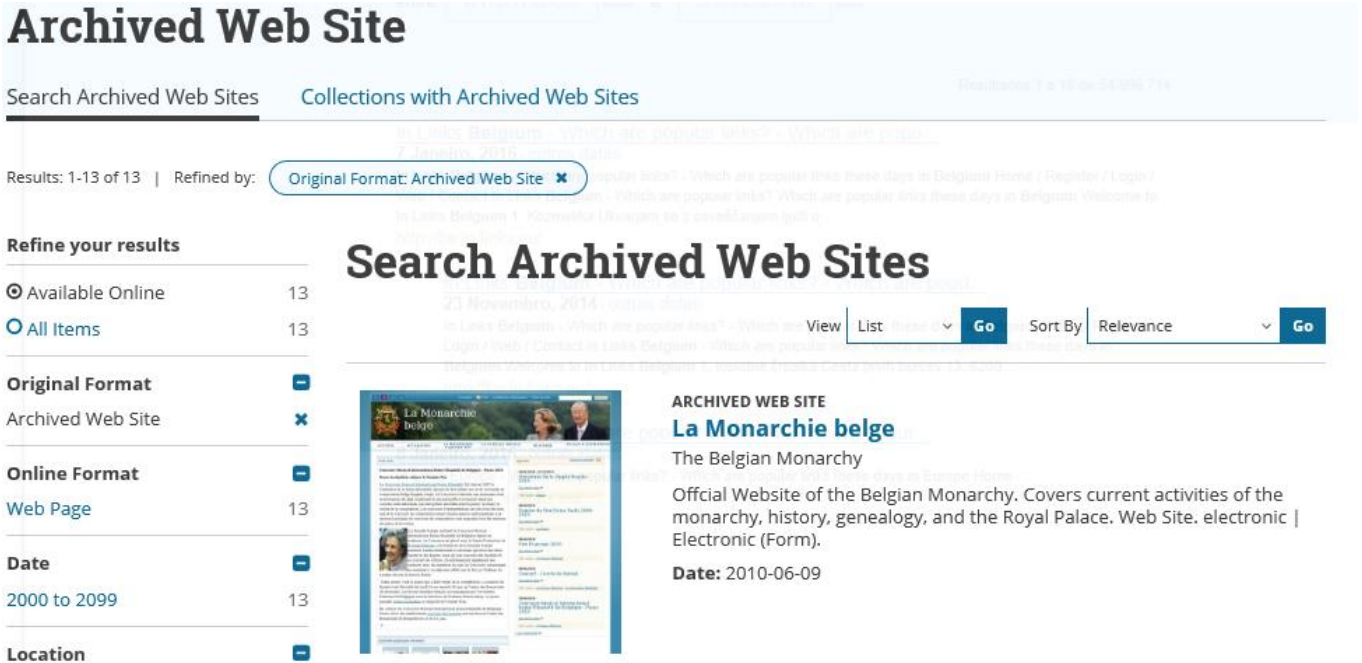
Summary of monarchie.be · Site Map of monarchie.be



promise

GHENT UNIVERSITY

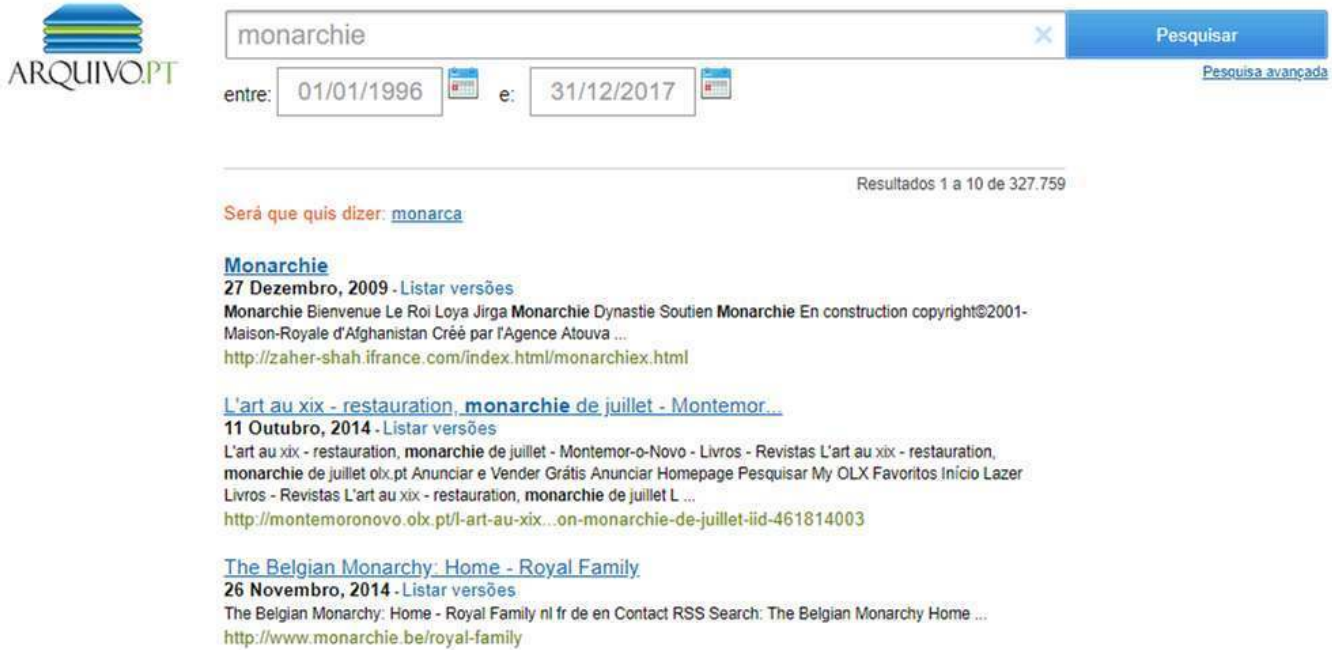
Library of Congress web archive



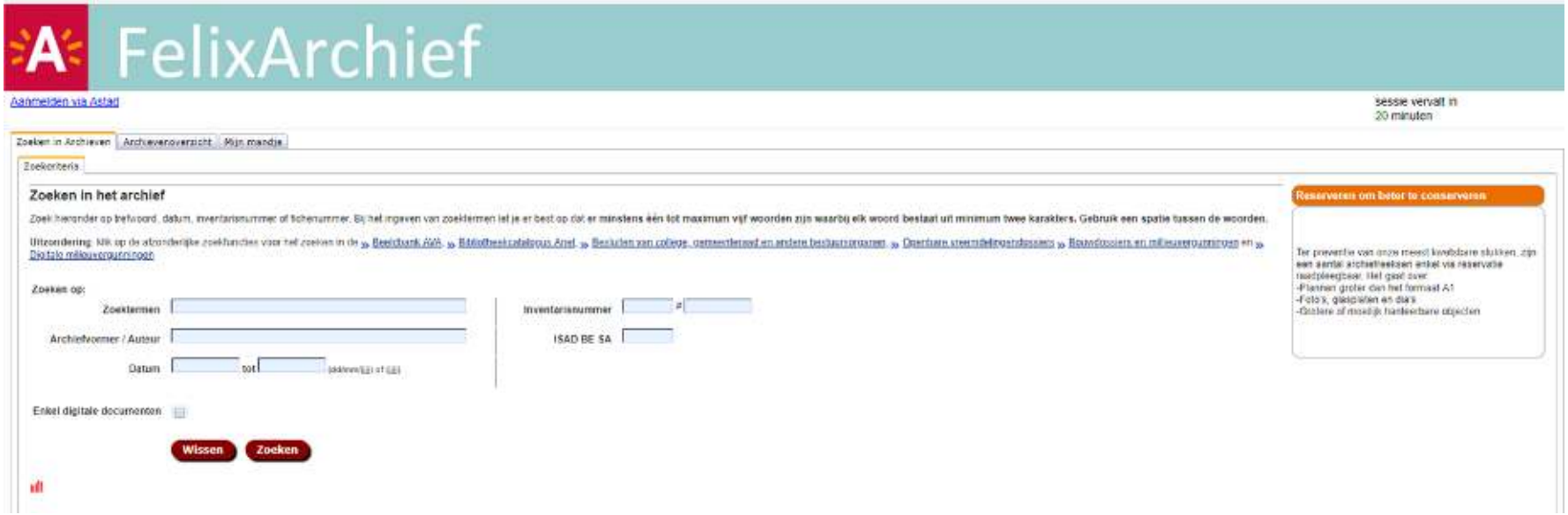
GHENT UNIVERSITY

Archives de l'État
Rijksarchief
www.arch.be

Arquivo.pt


ARQUIVO.PT
monarchie
Pesquisar
Pesquisa avançada
entre: 01/01/1996
e: 31/12/2017
Resultados 1 a 10 de 327.759
Será que quis dizer: monarca
Monarchie
27 Dezembro, 2009 - Listar versões
Monarchie Bienvenue Le Roi Loya Jirga Monarchie Dynastie Soutien Monarchie En construction copyright©2001-
Maison-Royale d'Afghanistan Créé par l'Agence Atouva ...
http://zaher-shah.ifrance.com/index.html/monarchiex.html
L'art au xix - restauration, monarchie de juillet - Montemor...
11 Outubro, 2014 - Listar versões
L'art au xix - restauration, monarchie de juillet - Montemor-o-Novo - Livros - Revistas L'art au xix - restauration,
monarchie de juillet olx.pt Anunciar e Vender Grátis Anunciar Homepage Pesquisar My OLX Favoritos Início Lazer
Livros - Revistas L'art au xix - restauration, monarchie de juillet L ...
http://montemoronovo.olx.pt/l-art-au-xix...on-monarchie-de-juillet-iid-461814003
The Belgian Monarchy: Home - Royal Family
26 Novembro, 2014 - Listar versões
The Belgian Monarchy: Home - Royal Family nl fr de en Contact RSS Search: The Belgian Monarchy Home ...
http://www.monarchie.be/royal-family

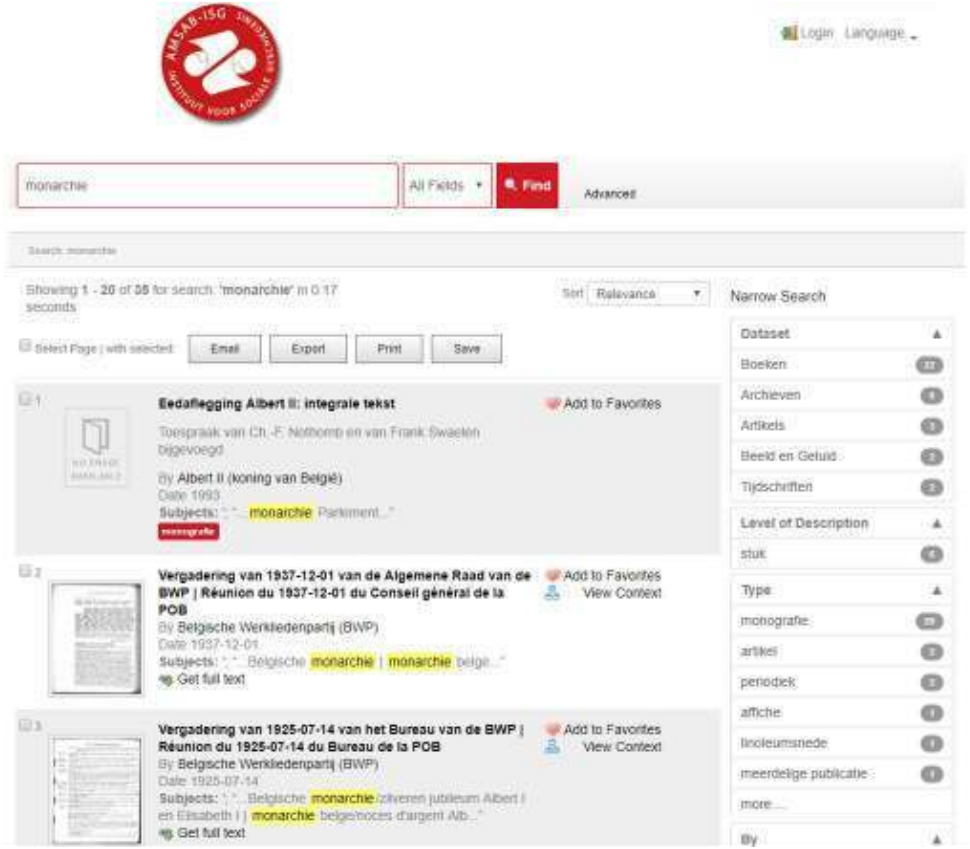
FelixArchief - Antwerp City Archives



GHENT UNIVERSITY

Archives de l'État
Rijksarchief
www.arch.be

AMSAB - Institute of Social Heritage



GHENT UNIVERSITY

Archives de l'État Rijksarchief
www.arch.be

Other topics

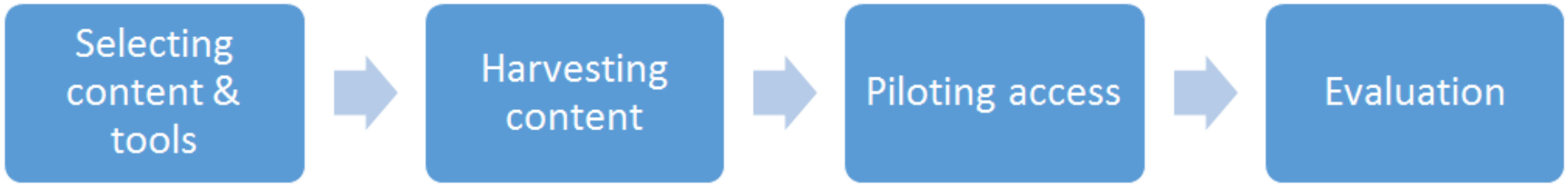
- Format of the web archived content
- Contribution to the selection of material to be included
- Important methods of analysis
- Challenges in working with data from web archives
- Subjects of most interest to be included
- Web archive functionalities
- Standardised question on “digital literacy”
- …

4. Next steps

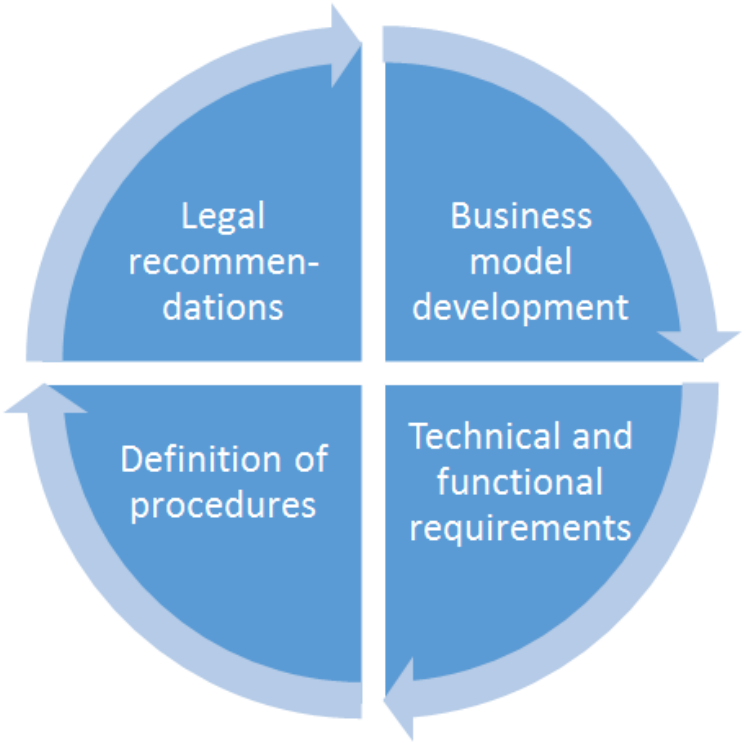
Phase 2: Belgian web archiving policy

- Analysis, topology and classification of the Belgian web
- Analysis of the legal framework for the Belgian web
- Proposal for a global Belgian web archiving strategy
- Selection criteria for appraisal methods
- Analysis of user requirement survey

Phase 3: Piloting web archiving



Phase 4: Recommendations for sustainable web archiving in Belgium



5. Q&A

Contact

Emmanuel Di Pretoro: edipretoro@he2b.be
Friedel Geeraert: friedel.geeraert@kbr.be
Alejandra Michel: alejandra.michel@unamur.be
Eveline Vlassenroot: eveline.vlassenroot@ugent.be