

Technical aspects

Gerald Haesendonck – IDLab UGent
Emmanuel Di Pretoro – HE²B



Scope

Initial scope of PROMISE

A prototype to ...

1. Select
2. Capture
3. Access

... web archives collections



Scope

Initial scope of PROMISE

A prototype to ...

1. Select
2. Capture
3. Access

... web archives collections

Needed during the project

1. Semi-automatic Quality Assessment (QA)
2. Derivative files to facilitate the use of the collections



Scope

Initial scope of PROMISE

A prototype to ...

1. Select
2. Capture
3. Access

... web archives collections

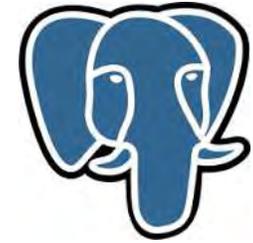
Needed during the project

1. Semi-automatic Quality Assessment (QA)
2. Derivative files to facilitate the use of the collections



1. Selection

In-house development (Python, Django, PostgreSQL)



django



KBR

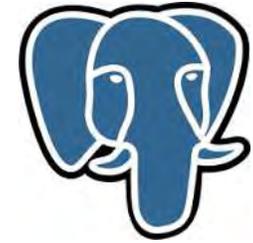


1. Selection

In-house development (Python, Django, PostgreSQL)

Basis for application

OCLC metadata set for web archives



django



KBR



1. Selection

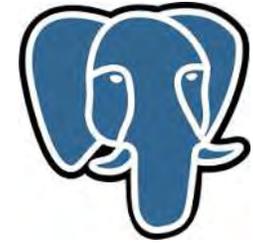
In-house development (Python, Django, PostgreSQL)

Basis for application

OCLC metadata set for web archives

First: simple tool to create seed lists

Later: automatically start crawling, trigger quality analysis, etc.



django



KBR



2. Capturing

Heritrix

broad crawls

configurable

fast

tried and tested

The logo for Heritrix, featuring the word "HERITRIX" in a bold, white, sans-serif font. The letters are set against a black background that has a jagged, blocky appearance, resembling a stylized cityscape or a digital interface.

2. Capturing

Heritrix

broad crawls

configurable

fast

tried and tested

The logo for Heritrix, featuring the word "HERITRIX" in a bold, white, sans-serif font. The letters are set against a black background that has a jagged, blocky appearance, resembling a stylized cityscape or a series of stacked blocks.

Browsertrix, Brozzler

high quality crawls

slower

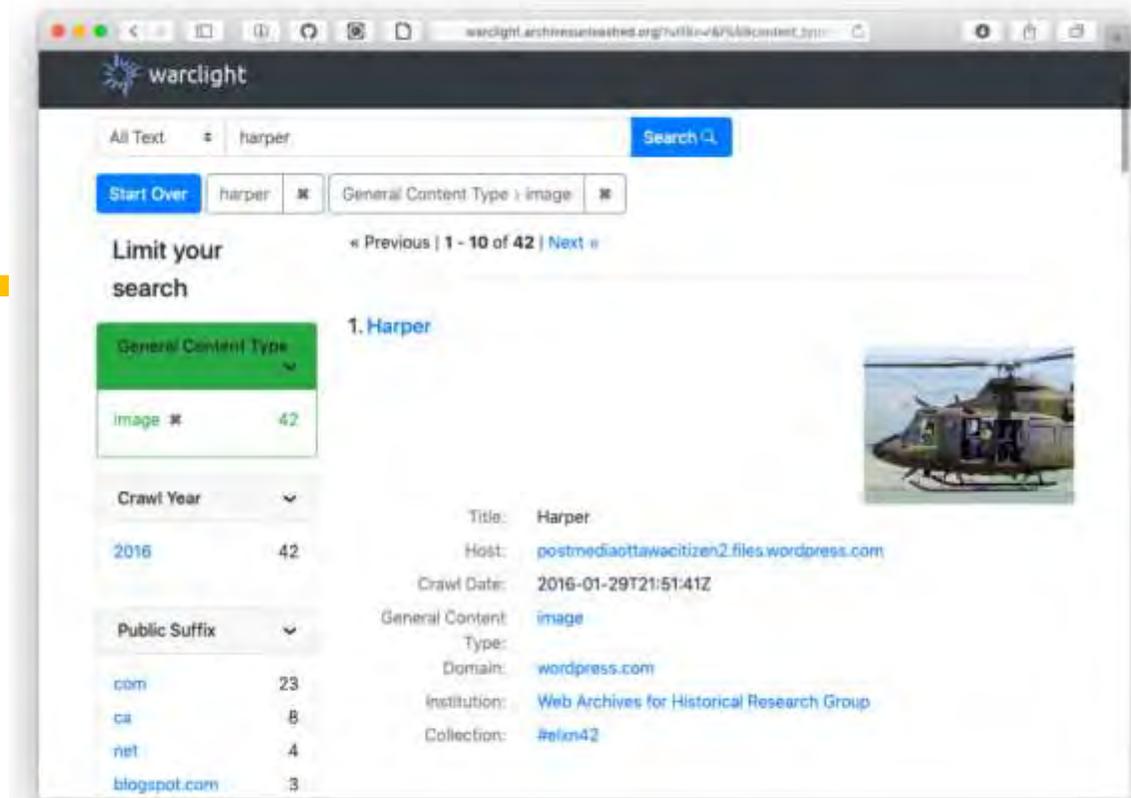
experimental



3. Access

Catalog

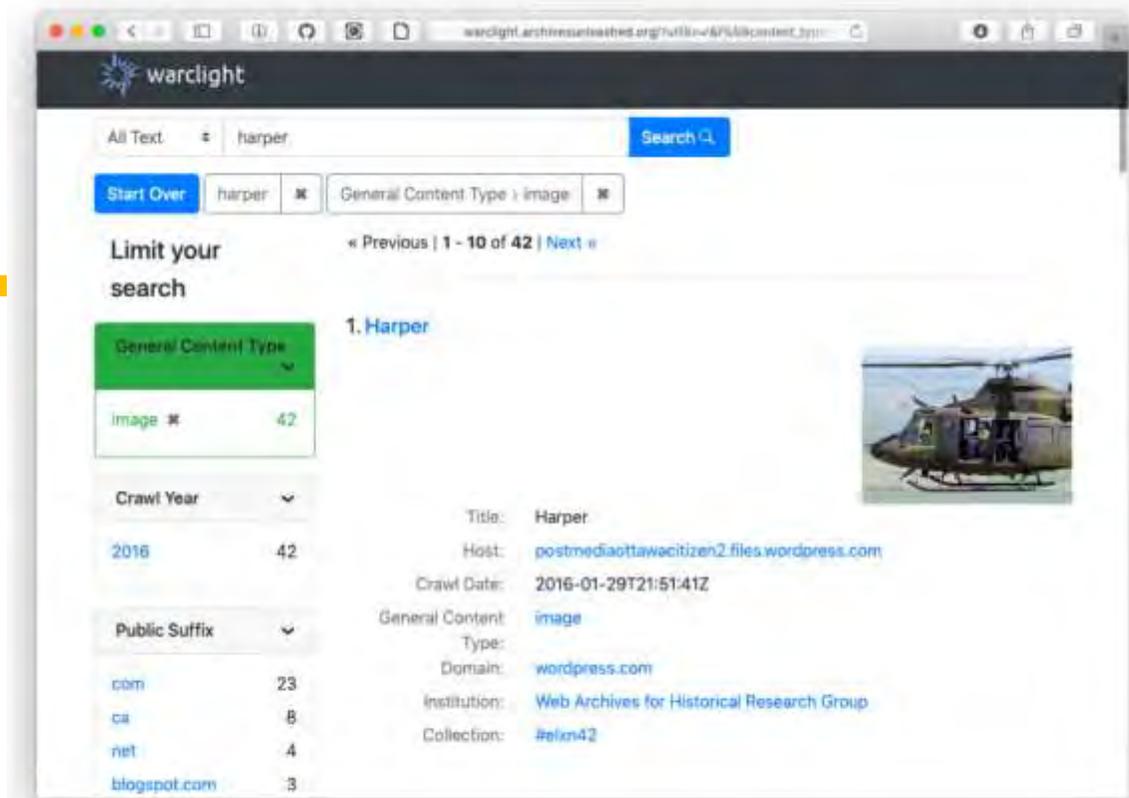
text search
discovery
based on WARCLight



3. Access

Catalog

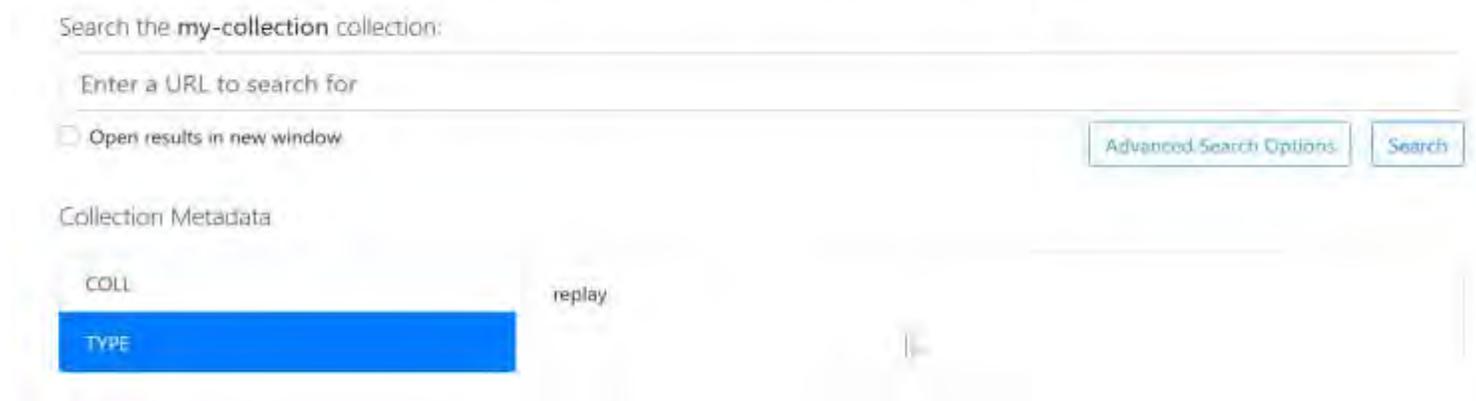
text search
discovery
based on WARCLight



Replay

URL search
timestamp
based on PyWB

Collection Search Page



Scope

Initial scope of PROMISE

A prototype to ...

1. Select
2. Capture
3. Access

... web archives collections

Needed during the project

1. Semi-automatic Quality Analysis
2. Derivative files to facilitate the use of the collections



Scope

Initial scope of PROMISE

A prototype to ...

1. Select
2. Capture
3. Access

... web archives collections

Needed during the project

1. Semi-automatic Quality Analysis
2. Derivative files to facilitate the use of the collections



Web Archive Quality Analysis

How well can we capture and archive web content?
Can we check this **semi-automatically**?



Quality of content

Visual correspondence
“Does it **look** the same?”

Reyes, Brenda. A Grounded Theory of Information Quality in Web Archives, dissertation, August 2018; Denton, Texas.



Quality of content

Visual correspondence

“Does it **look** the same?”

Interactional correspondence (IC)

“Can you **interact** the same way?”

Reyes, Brenda. A Grounded Theory of Information Quality in Web Archives, dissertation, August 2018; Denton, Texas.



Quality of content

Visual correspondence

“Does it **look** the same?”

Interactional correspondence (IC)

“Can you **interact** the same way?”

Completeness

“Do we have **every resource** of the original?”

Reyes, Brenda. A Grounded Theory of Information Quality in Web Archives, dissertation, August 2018; Denton, Texas.



Quality of content

Visual correspondence

“Does it **look** the same?”

Interactional correspondence (IC)

“Can you **interact** the same way?”

Completeness

“Do we have **every resource** of the original?”

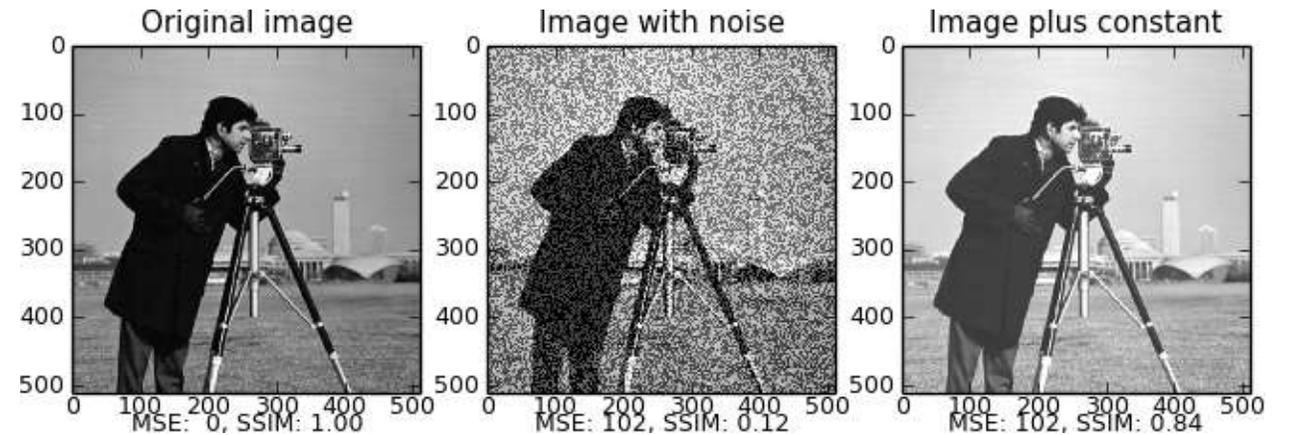
Reyes, Brenda. A Grounded Theory of Information Quality in Web Archives, dissertation, August 2018; Denton, Texas.



Visual correspondence

Quality metrics based on:

Structural Similarity (**SSIM**)
sensitive to noise
less sensitive to colour



Z. Wang et Al. Image quality assessment: from error visibility to structural similarity. IEEE Transactions on Image Processing

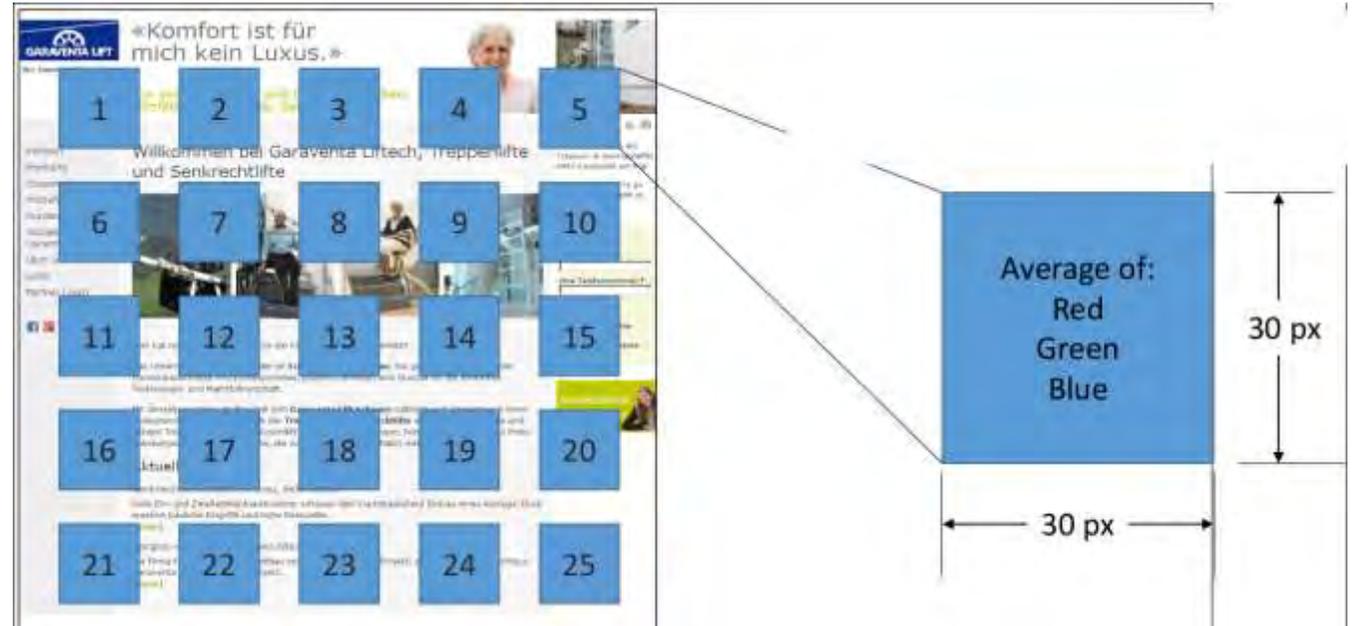
Image: Copyright the scikit-image development team



Visual correspondence

Quality metrics based on:

Visual Quality Indicator (VQI)
sensitive to colour
less sensitive to noise



Swiss National Library. Visual Quality Indicator

Image: Copyright Swiss National Library



Automated Quality Metrics

The SSIM of <https://www.bedetheque.com/auteur-2872-BD-Van-Hasselt-Thierry.html> is of **25.05%**.

The VQI of <https://www.bedetheque.com/auteur-2872-BD-Van-Hasselt-Thierry.html> is of **13623.93**. That means that the archived version is **not acceptable**.

Visualization of the differences

Original version with marks

This screenshot shows the original website with several yellow highlights and red marks. At the top, there is a banner for 'COLLECTIONNEZ LES VOITURES DE TINTIN' with a '1/24*' scale and '1 MODELE ACHETE = 1 MODELE OFFERT + DES CADEAUX!' offer. Below this, the main content area is highlighted in yellow, showing the author's profile for 'Van Hasselt, Thierry'. The profile includes a bio, a list of works, and a table of publications. The table has columns for 'Date de parution', 'Titre', and 'Statut'. The bottom of the page features a navigation menu with categories like 'A la Une', 'Actualités', 'La légende', 'Communauté', and 'Bibliothèque'. The page is framed by a yellow border with illustrations of vintage cars and the 'Le Monde hachette' logo.

Archived version with marks

This screenshot shows the archived version of the website. The page is mostly black, with the original content visible in white. The top banner and main content area are highlighted in black. The author's profile for 'Van Hasselt, Thierry' is visible, but the bio and list of works are mostly obscured by black boxes. The table of publications is also mostly black. The bottom navigation menu is visible but mostly obscured. The page is framed by a black border with illustrations of vintage cars and the 'Le Monde hachette' logo. The overall quality is significantly lower than the original version.

Automated Quality Metrics

The SSIM of <https://directory.unamur.be/teaching/programmes/050K> is of **89.09%**.

The VQI of <https://directory.unamur.be/teaching/programmes/050K> is of **33.83**. That means that the archived version is **acceptable**.

Visualization of the differences

Original version with marks

This screenshot shows the original website interface for 'ANNUAIRE DES FORMATIONS' at the University of Namur. The page features a search bar at the top, a navigation menu, and a main content area with a title 'Master de spécialisation en conservation-restauration du patrimoine culturel immobilier'. The page is annotated with numerous red bounding boxes that highlight specific elements, including the university logo, navigation links, the main title, and various text blocks and buttons throughout the page.

Archived version with marks

This screenshot shows the archived version of the website. The top navigation bar is dark grey and includes the text 'U Namur - Annuaire des formations' and 'Archivé en Wed, 23 Jan 2019 10:43:13 GMT'. The main content area is identical to the original version, but the page is annotated with red bounding boxes that highlight differences from the original. These differences are primarily located in the navigation menu, the main title, and various text blocks and buttons throughout the page.

Quality of content

Visual correspondence

“Does it **look** the same?”

Interactional correspondence (IC)

“Can you **interact** the same way?”

Completeness

“Do we have **every resource** of the original?”

Reyes, Brenda. A Grounded Theory of Information Quality in Web Archives, dissertation, August 2018; Denton, Texas.



Interactional correspondence (IC)

Degree to which a user's interaction with the archived website is similar to that of the original



Interactional correspondence (IC)

Degree to which a user's interaction with the archived website is similar to that of the original

Idea:

- interacting (e.g.: clicking link) results in browser **requests** (HTML, images, CSS, JS, ...).
- How much of these requests are **successful in archive?**



KBR



IC

$$= \frac{\# \text{ successful requests in archived website}}{\# \text{ requests in original website}}$$



KBR



IC

$$= \frac{\# \text{ successful requests in archived website}}{\# \text{ requests in original website}}$$

Requests can be **weighted** by their **importance** to decrease the impact of less important requests



IC: our approach

1. Build **index** from archive → fast lookups



IC: our approach

1. Build **index** from archive → fast lookups
2. Crawl page from archive, **capture all requests**

Optionally filter out ads
no impact on user interaction



IC: our approach

1. Build **index** from archive → fast lookups
2. Crawl page from archive, **capture all requests**

Optionally filter out ads
no impact on user interaction

3. Determine resource importance



IC: resource importance

Important factors:

- Content type: HTML > images > fonts
- CSS coverage
- Image size & position

J.F. Brunelle et Al. Not all mementos are created equal: measuring the impact of missing resources. International Journal on Digital Libraries, September 2015, Volume 16, Issue 3-4, pp 283-301



IC: resource importance

Important factors:

- Content type: HTML > images > fonts

- CSS coverage

- Image size & position

J.F. Brunelle et Al. Not all mementos are created equal: measuring the impact of missing resources. International Journal on Digital Libraries, September 2015, Volume 16, Issue 3-4, pp 283-301



(a)



A Stylesheet is important when:

- Disabling makes content shift left



A Stylesheet is important when:

- Disabling makes content shift left



KBR



A Stylesheet is important when:

- Disabling makes content shift left
- It has high coverage



KBR



A Stylesheet is important when:

- Disabling makes content shift left
- It has high coverage

An image is important when it:

- Is large
- Overlaps the horizontal center
- Is in the 70% vertical center



KBR



Scope

Initial scope of PROMISE

A prototype to ...

1. Select
2. Capture
3. Access

... web archives collections

Needed during the project

1. Semi-automatic Quality Analysis
2. Derivative files to facilitate the use of the collections



Web Archive Derivatives

Represent (a part of) the archive in a way suited to **answer certain questions.**



Why?

Provide information **about** a web archive



Why?

Provide information **about** a web archive

Facilitate **analysis and research** on the archive



Why?

Provide information **about** a web archive

Facilitate **analysis and research** on the archive

Often **smaller** in size, more efficient to process



Chosen derivatives

Archive-It: **WAT**, WANE, LGA

Metadata on record level

- URL
- Original archive file name
- Document information. E.g. HTML:
 - Title
 - Keywords
 - Scripts
 - Links



Chosen derivatives

Archive-It: WAT, **WANE**, LGA

Named entities

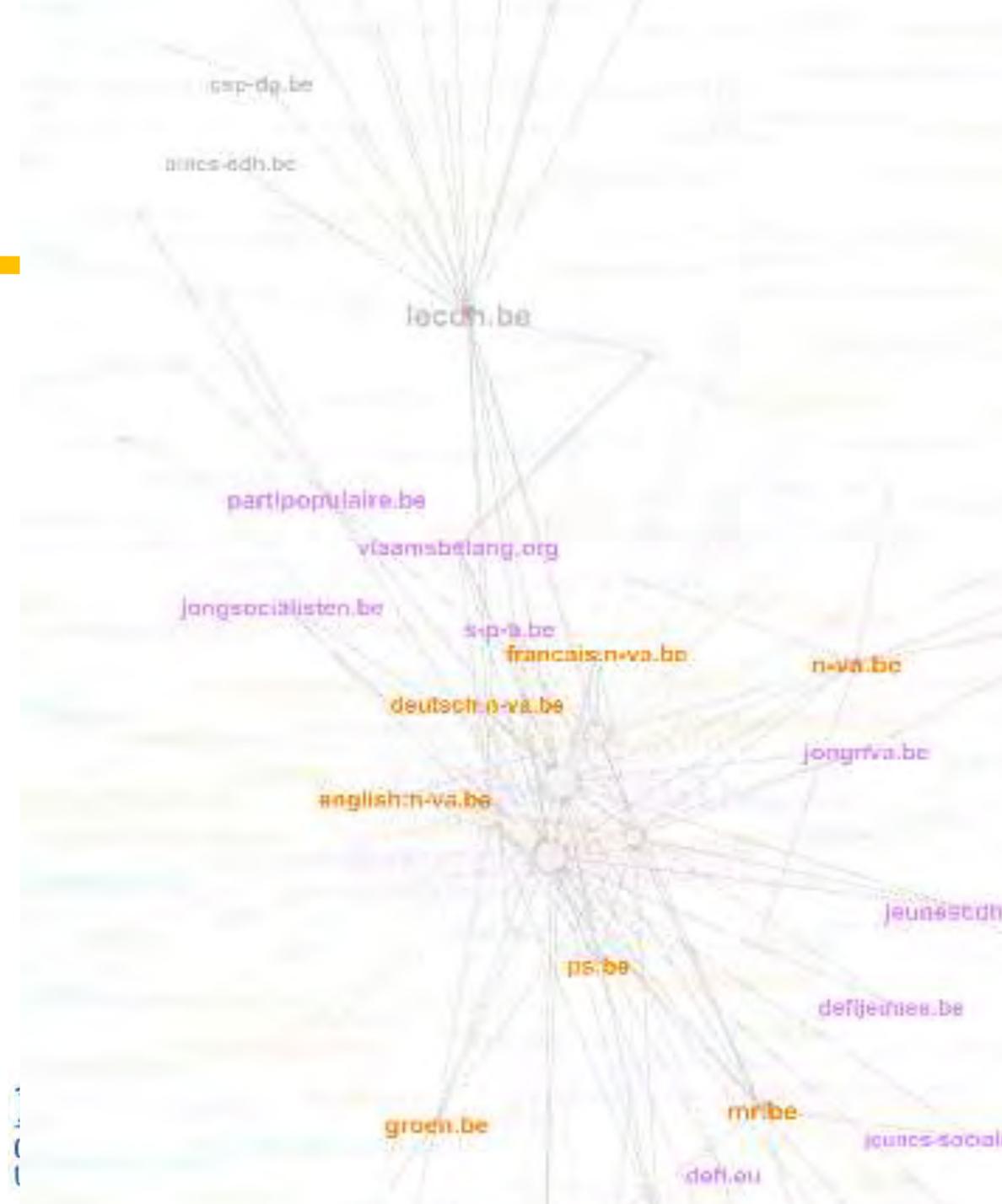
```
{"url": "https://opac.kbr.be/about.aspx?_lg=en-GB",  
  "named_entities": {  
    "persons":  
      ["Archimed"],  
    "organizations":  
      ["Royal Library", "OPAC", "KBR"],  
    "locations":  
      ["Brussels", "Belgium"]  
  }  
}
```



Chosen derivatives

Archive-It: WAT, WANE, LGA

Graph of web pages



KBR



Chosen derivatives

Archives Unleashed Toolkit: **Domains**, Plain text, GraphML

Domain occurrences in archive

(opac.kbr.be, 7972)	(drp.kbr.be, 14)
(belgica.kbr.be, 6427)	(bartok.kbr.be, 11)
(isil.kbr.be, 2746)	(193.190.242.40, 10)
(www.kbr.be, 1766)	(coins.kbr.be, 10)
(uur1.kbr.be, 807)	(www.w3.org, 9)
(events.kbr.be, 286)	(kbr.prezly.com, 8)
(www.depotlegal.be, 250)	(www.ultimedia.com, 6)
(ysaye.kbr.be, 212)	(www.google.com, 6)
(vieuxtemps.kbr.be, 129)	(lms-web-srv01.kbr.be, 4)
(www.adobe.com, 64)	(maps.googleapis.com, 4)
(www.ngi.be, 28)	(sharethis.com, 4)
(www.youtube.com, 16)	(get.adobe.com, 3)



Chosen derivatives

Archives Unleashed Toolkit: Domains, **Plain text**, GraphML

Web documents as plain text

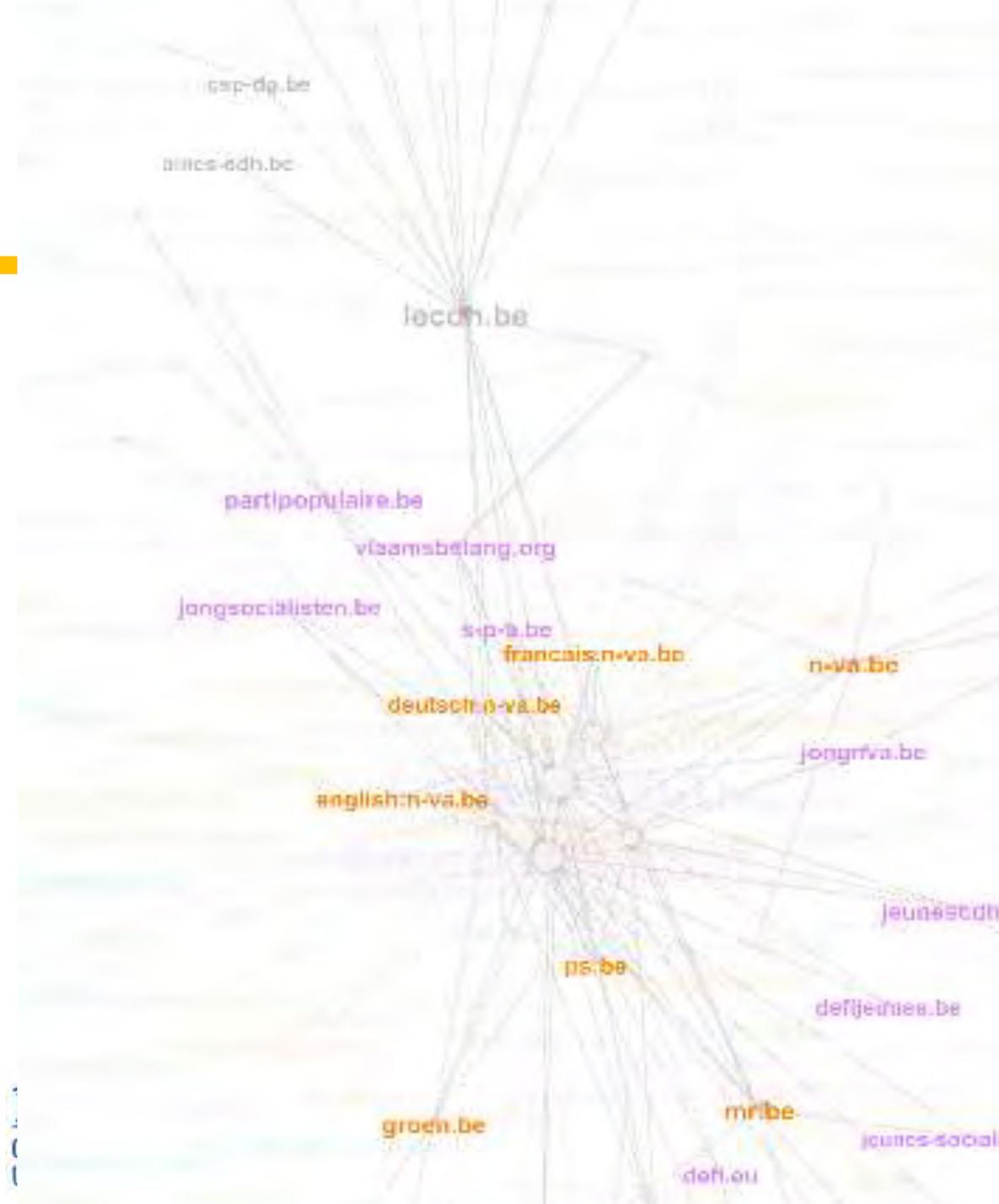
(20190312,opac.kbr.be,https://opac.kbr.be/library/about.aspx?_lg=en-GB,OPAC - About the catalog | Royal Library of Belgium Show menu EN NL FR EN OPAC OPAC Other sites Bibliothèque fédérale Dépôt légal Educatif Galerie My account My account Go to menu Go to content Go to search kbr.be OPAC Bibliothèque fédérale Dépôt légal Educatif Galerie My accountMy account Syracuse media library Your login ID Your login ID Your password Your password OK Register OK Register Catalogue, selected Catalogue Catalogue Search input field Clear search field Start search on the script Advanced search You are here: Home / About the catalog | Royal Library of Belgium / Item details Pré-sélectionner des critères de recherche Modifier les critères pré-sélectionnés NL FR EN FAQ About the online catalog of the KBR Welcome to the online catalog of the Royal Library of Belgium. What you can and cannot find in this catalog The majority of the collection of the Royal Library is recorded in this catalog, but certain documents aren't (yet) : 1. A lot of manuscripts aren't recorded in the catalog because they are difficult to classify according to modern standards.



Chosen derivatives

Archives Unleashed Toolkit:
Domains, Plain text, **GraphML**

Graph, similar to LGA



KBR



Chosen derivatives

In-house: web page extraction

Extract data from certain domain(s) or web pages

Researcher can work on raw data with small size



Chosen derivatives

In-house: web page extraction

Extract data from certain domain(s) or web pages

Researcher can work on raw data with small size

Approach:

- Build index on archive → fast lookup
- Crawl the archive
- Put output in new archive



Technical aspects

Gerald Haesendonck – IDLab UGent
Emmanuel Di Pretoro – HE²B

