

# The potential of the archived web for scholarly use

---

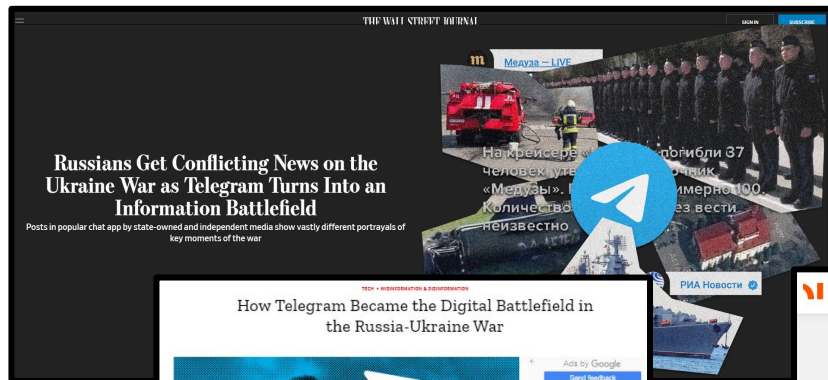
Presentation #1: Theoretical frameworks, research questions and axiological positions

7 september 2022, RedMIL Summer School  
Eveline Vlassenroot, Ghent University

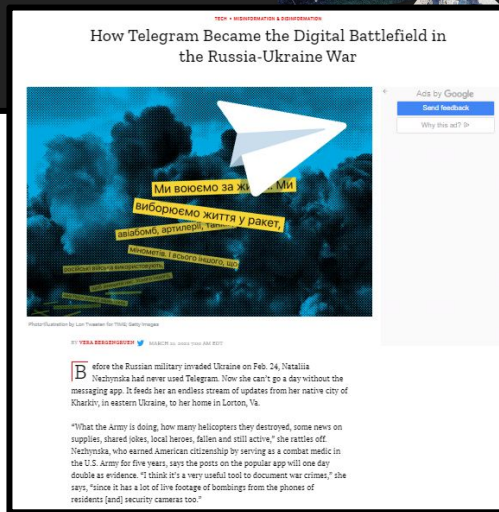
# Content

1. Problem statement
2. Research question(s)
3. Conceptual framework
4. Axiological position
5. Expected findings

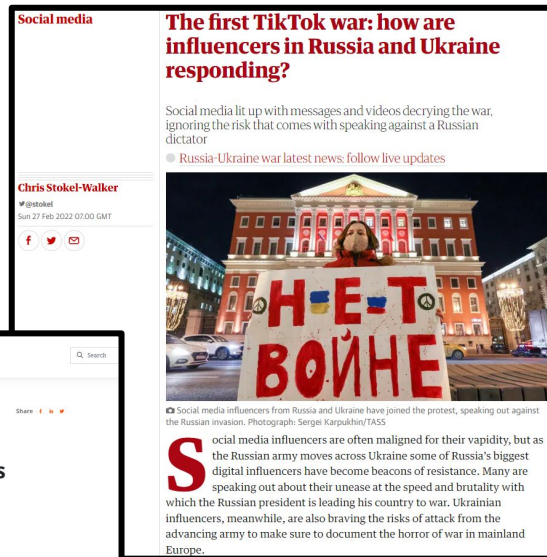
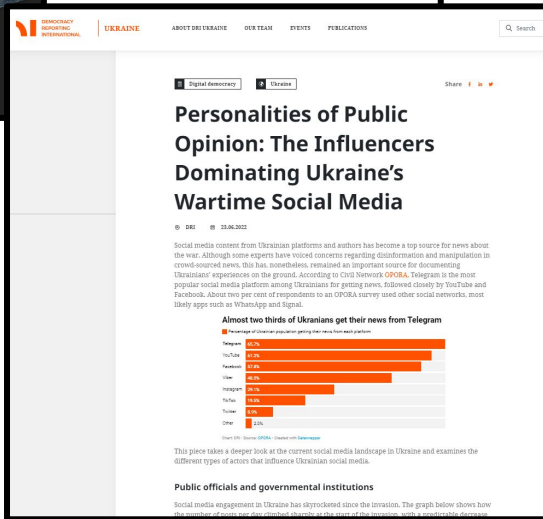
# The Web as an important footprint of the past



Source:  
<https://www.wsj.com/articles/russians-get-conflicting-news-on-the-ukraine-war-as-telegram-turns-into-an-information-battlefield-11654314693>



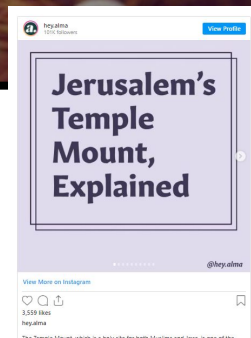
Source:  
<https://time.com/6158437/telegram-russia-ukraine-information-war/>



Source:  
<https://www.theguardian.com/media/2022/feb/26/social-media-influencers-russia-ukraine-tiktok-instagram>

Source:  
<https://democracy-reporting.org/en/office/ukraine/publications/personalities-of-public-opinion-the-influencers-dominating-ukraines-wartime-social-media>

# Relevance for researchers



META NUMMER 2021/4

## Gespreksavond dekolonisatie in actie: from tweets to the streets

GESCHREVEN DOOR  
[Jeroen Fernandez-Alonso](#)

GEPUBLICEERD OP  
11.05.2021

DEEL DEZE PAGINA  
f t e p



## Hoe archiveren we het online luik van het activisme van de 21ste eeuw?

Amsab-Instituut voor Sociale Geschiedenis (Amsab-ISG) organiseerde op 12 maart een [gespreksavond](#) met als thema's: dekolonisatie, de rol van participatie op sociale media, en hoe we dat het best kunnen bewaren voor de toekomst. De gespreksavond kaderde binnen het Gentse Belmundofestival. Dat is een jaarlijks activiteitenfestival rond internationale solidariteit waarbij de Stad Gent en 11.11.11 de handen in elkaar slaan met het Gentse middenveld, en waar Amsab-ISG ook jaarlijks aan deelneemt.

# International web archives since 1996 ...

1996

Australia: PANDORA

UK Government web  
archive

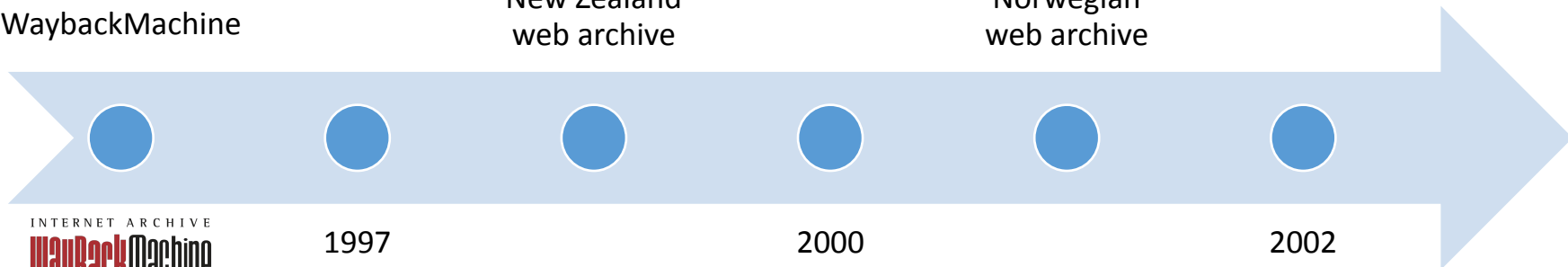
Internet Archive:  
WaybackMachine

1999

New Zealand  
web archive

2001

Norwegian  
web archive



INTERNET ARCHIVE  
**WayBackMachine**

1997

Sweden:  
Kulturarw3

2000

USA: Library of Congress  
web archive

Czech Republic:  
Webarchiv

2002

France: BnF  
web archive

# ... but what about Belgium?



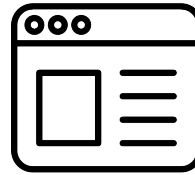
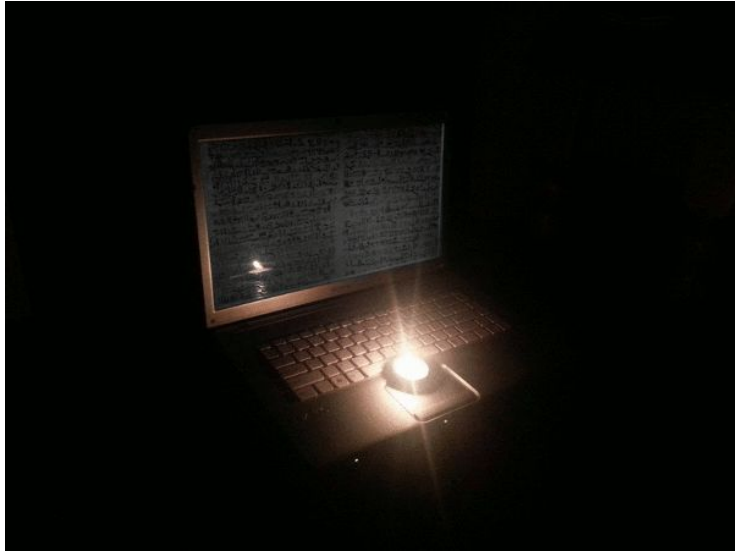
2017 – 2019  
Focus on  
archiving websites



2020-2022  
Focus on  
social media  
archiving

Upcoming project: 2023-2025 BelgicaWeb

# Challenge 1: Short lifespan of online data

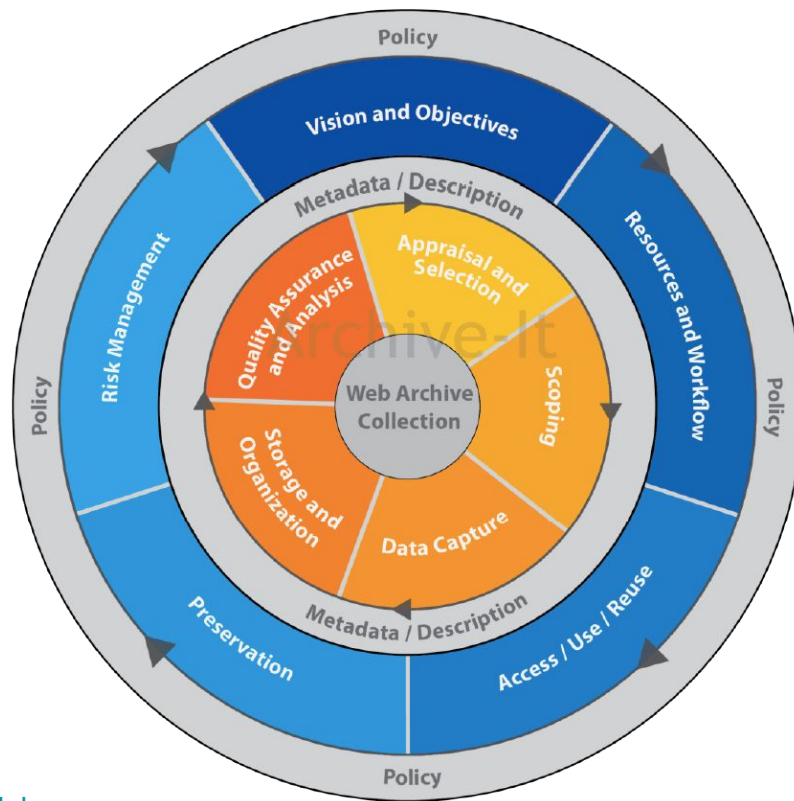


Icons made by Vectors Market & Pixel Market from  
[www.flaticon.com](http://www.flaticon.com)

# Challenge 2: Appropriate data analysis



# Challenge 3: Development of web archives



Source: [Web Archive IT Lifecycle Model](#)

# Research question

RQ: What is the potential of web archived content for scholarly use?

RQ1: **What knowledge, values and skills are needed for researchers to work with web archives?**

# What is “web archiving”?

Web archiving is the **process of gathering up data** that has been **recorded on the World Wide Web**, **storing it**, ensuring the data is **preserved** in an archive, and making the collected data **available** for future research. (Niu, 2012)

The web archive as **a form of “media”** as it channels cultural productions (collections). Like other forms of media, technical infrastructure, design, funding, role expectations, and distribution methods shape the digital archive. (Jensen, 2021)

# Objective of this research

- 1) To design and validate the Web Archival Intelligence Scale (WAIS) as an instrument to **measure knowledge, values and skills** of researchers to work with web archived content.
- 2) To provide archiving institutions with **guidance on methods to use in their own services** that offer web archived content.

# Web Archival Intelligence Scale (WAIS)

- Related to digital and archival literacy
- To allow the **technical and cognitive understanding** of the archives (this requires comprehension of both their shaping and contextualisation) (Bachimont, 2017)

# Theoretical framework: “Archival Literacy”

Instrument based on 3 types of knowledge, values and skills (Yakel & Thores, 2003):

1) Subject (or domain) knowledge (Winter, 1993)

*“An understanding of the topic being researched, for example historic preservation, politics, or the law.”*

2) Artifactual literacy (Bass, 2002)

*“ The practice of criticism, analysis, and pedagogy that reads texts as if they were objects and objects as if they were texts.”*

3) Archival Intelligence (Yakel & Thores, 2003)

*“A researcher’s knowledge of archival principles, practices, and institutions, such as the reasons underlying archival rules and procedures, how to develop search strategies to explore research questions, and an understanding of the relationship between primary sources and their surrogates.”*

# Challenges of using web archived content (1)

## Methodological:

- Impossible to return to the original document
- The data that was collected is probably incomplete
- Inconsistent in time&space, not everything was archived at the same time + not at the same depth
- What goes into the web archive, is not the web as seen in a web browser

# Challenges of using web archived content (2)

## Technical:

- The use of the format “WARC” - which is not known outside of the community
- Size of the files - web archives can be quite large
- Processing of the WARC files - lots of dependencies exist

## Conceptual:

- Provenance - context is needed of why the collections take the form as they do
- Acquisition - many different ways exist to acquire and harvest content
- Borders on a crawl - these can be very fluid
- Size of the data - the more data, the more challenging the data derivation
- Gaps in collections - both intentional or unintentional

# Challenges of using web archived content (3)

## Practical:

- Unfamiliarity with web archives
- When making a request for data, does the researcher know who he is talking to?
- Agreements for sharing data

# Axiological position

To whom, outside of academia, is your research meant to be useful or relevant?

- Institutions constructing web archives
- The web archiving community
- Teachers in multiple disciplines

# Expected contributions

## 1) Academic contribution:

By studying the growing potential of web archives, raising awareness of the existence of web archives and identifying the needs and requirements of researchers to use these web archives.

## 2) Methodological contribution:

The development of the Web Archival Intelligence Scale (WAIS) as contribution towards the domain of data and archival literacy.

## 3) Practical contribution



Policy contributions towards web archiving institutions in the format of mock-up interfaces and decision trees as an example to guide researchers through the web archiving institution's interfaces to find relevant material for their research.

How can the existing literature on literacies & competences help the development of the Web Archival Intelligence Scale?

# Thank you!

---

Eveline.Vlassenroot@UGent.be

# The potential of the archived web for scholarly use

---

Presentation #2: Research methods and epistemology

9 september 2022, RedMIL Summer School  
Eveline Vlassenroot, Ghent University

# Research questions

General RQ: What is the potential of web archived content for scholarly use?

RQ1: What are the needs and requirements for (humanities) researchers to work with web archives?

RQ2: What *knowledge, values and skills* are needed for (humanities) researcher to work with web archives?

## **RQ1: What are the needs and requirements for (humanities) researchers to work with web archives?**

---

Research in three phases:

- 1) Desk research
- 2) Semi-structured interviews (n=6)
- 3) Pre-conference workshop @RESAW

## **RQ1: What are the needs and requirements for (humanities) researchers to work with web archives?**

---

### **1) Desk research**

- User requirements studies by Costea (2018); Riley & Crookston (2015); BnF (2011), Costa & Silva (2010); Dougherty et al. (2010); etc.
- “Orientating”, “auditing” and “constructing” as three conceptual devices to describe common research practices for researchers engaging with web archives (Ogden & Maemura, 2021)

## RQ1: What are the needs and requirements for (humanities) researchers to work with web archives?

---

### 2) Semi-structured interviews (1/3)

- Allow for **more flexibility** in which topic lists do not need to be followed religiously and can be modified depending on the expertise or the issues raised during the conversation
- Approach a problem **from within the context of the research subject** and reveal **both factual** knowledge as well as the **opinions** of the interviewees
- Recruitment: via **persona** outlines
  - *“A persona is a fictional character created to represent a user type that might use a social media archive in a similar way.” (Pruitt et al., 2003)*

# RQ1: What are the needs and requirements for (humanities) researchers to work with web archives?

## 2) Semi-structured interviews (2/3)

Five personas are described in order to help with “(...) *l’identification et à la définition des profils des usagers potentiels*” (Moiraghi, 2018, p. 38).



Fig. 1: Overview personas Corpus-project @BnF



- **Bart** - post-doc researcher in communication sciences



- **Febe** - PhD-student in computational linguistics



- **Ben** - 35-year old journalist working for the newspaper



- **Manou** - 28-year old scientific researcher @ KBR



- **Jan** - 68-year old former team coordinator

Fig. 2: Overview personas BeSocial project @KBR

**RQ1: What are the needs and requirements for (humanities) researchers to work with web archives?**

---

## **2) Semi-structured interviews (3/3)**

- Using these **5 personas** as an “ideal” we explored our personal and professional networks to find suitable interviewees and contacted them using email.
- **6 interviews** were conducted in 2021
- Interview guide (using both pre-defined and open questions) encompassed **3 main themes**: selection, consultation and involvement & support

# RQ1: What are the needs and requirements for (humanities) researchers to work with web archives?

---

## 3) Pre-conference workshop @RESAW (n=15)

*“Title: User requirements for working with social media web archives”*

Workshop structure:

- 1) General introduction & context;
- 2) Four interesting speakers shared insights on what it means to preserve (social) media content and on what it takes to use (social) media content as research data
- 3) **Hands-on workshop** via break-out rooms: Users pick a “BeSocial-Person” and decide in group on a web-archive use case and reflect on it from their persona for three separate phases (orientating/auditing/constructing)



Fig. 3: Overview of some of the workshop results on ‘orientating’

## **RQ2: What *knowledge, values and skills* are needed for (humanities) researcher to work with web archives?**

---

=> How can I help social media archives in serving their users with the right tools, functionalities etc?

I want to design an instrument/scale to **measure the competencies and needs of the users** by following a multi-phase scale development approach (DeVellis, 2012):

**Phase 1:** Item generation and scale development phase

**Phase 2:** Testing phase

- Exploratory & confirmatory factor analysis
- Item convergent validity & internal consistency

## RQ2: What *knowledge, values and skills* are needed for (humanities) researcher to work with web archives?

---

### 1) Item generation and scale development



|        |                           |                                         |                                                                                                                                                                                             |
|--------|---------------------------|-----------------------------------------|---------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| Step 1 | Literature review         | Inventory of XX items                   | Searching for studies describing forms of literacies, skills and competences for researchers working with (web) archived content and how these modes conceptualized, described and measured |
| Step 2 | Expert panel review 1     | Preliminary version of scale (XX items) | Reformulation of the items to better fit into the context and technological affordances of web archives                                                                                     |
| Step 3 | Expert panel review 2     | Qualitative assessment of scale         | Evaluation the different items of the scale in terms of wording, grammar, scaling and item allocation                                                                                       |
| Step 4 | *New* expert panel review | Quantitative assessment of scale        | Rating of the items in terms of relevance, clarity and simplicity                                                                                                                           |

**RQ2: What *knowledge, values and skills* are needed for (humanities) researcher to work with web archives?**

---

## **2) Testing**

Development of a **questionnaire** (or include in an existing survey):

- Include the items of WAIS (see phase 1)
- Collect socio-demographic characteristics

**Statistical analysis** to test the psychometric properties of WAIS:

- Construct validity: exploratory and confirmatory factor analysis
- Item convergent validity
- Internal consistency (reliability)

# Thank you!

---

Eveline.Vlassenroot@UGent.be