AARHUS
UNIVERSITY

3 NOVEMBER 2015

Web History, Web archives, and Web Research Infrastructure — between close and distant reading

Niels Brügger
PROFESSOR,
HEAD OF THE CENTRE FOR
INTERNET STUDIES, AND OF NETLAB

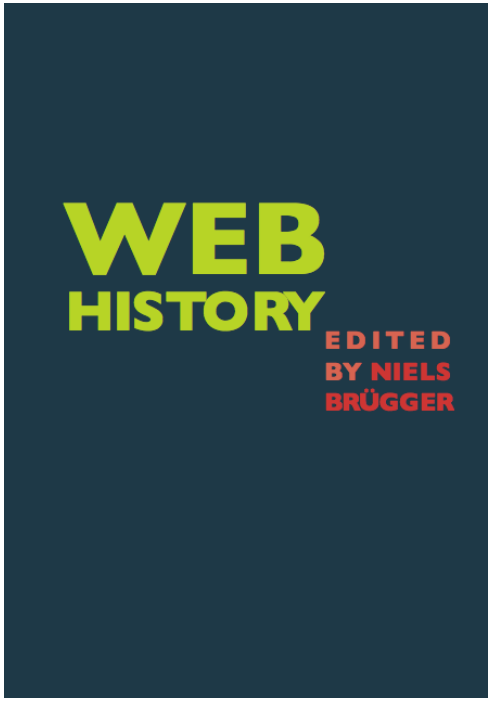
UNIVERSITET

AGENDA

1. Web history, web archives — a personal account
2. Web history — Digital history
3. General characteristics of archived web
4. The scholarly use of web archives
5. Scholarly needs? A matrix for a generic approach
6. The archived web used as historical source
7. The history of dr.dk 1995-2005
8. The Danish Web 2005-2015
9. Analysis of big archived web corpora: challenges?
10. What is NetLab and RESAW?

1. Web history, web archives — a personal account

Web historian — historical interest in the web started in the late 1990s



2010 (Peter Lang)

1. Web history, web archives — a personal account

What I was missing in the late 1990s:

1. A preserved version of my object of study, the web of the past
2. Methodological and theoretical reflections about the scholarly use of archived web
3. A research infrastructure and analytical tools

1. Web history, web archives — a personal account

From a personal solution to a national web archive

INTERNETAVISEN
Jyllands-Posten
Torsdag 7. september

| Om JP | Job | Abonnent | Arkiv | Morgenavisen | Internetavisen |

INTERNETAVISEN
Forside
Indland
Udland
Sport
Erhverv
Børs & Finans
IT & Computer

JP København
JP Århus

Karriere & Privatøkonomi
JP Explorer
JP Guide
Vejret

ANNONCER

test dig selv - klik her

Nedtur i hårdt kvarter er vendt

Indland: En målrettet indsats gav resultat i det belastede boligområde Avedøre Stationsby. På kort tid er kriminaliteten og arbejdsløsheden blevet halveret. Læs artikel

Thorsen-sagen først i retten til næste sommer

Erhverv: Ankesagen mod tidligere byggematador Kurt Thorsen og Rasmus Trads kommer først for i Østre Landsret til næste sommer. Læs artikel

JP-Xtra ◆ Bred opbakning til hårde PFA-domme ◆ Direkte TV fra PFA-dom ◆ Injuriesag på vej i Thorsen-sagen

Telia vil overtage Tele Danmarks mobilvirksomhed

Børs & Finans: Telia ønsker at overtage Tele Danmarks mobilvirksomhed. Det siger administrerende direktør Kenneth Karlberg, Telia Mobile. Han vurderer, at en aftale vil være 50 mia. kr. værd. Læs artikel

Mobiltelefonerne gør de unge til en del af et socialt kredsløb. Foto: Lizette Kabré

Børn uden mobiltelefon er på socialt sidespor

Indland: Børn uden en mobiltelefon risikerer at komme ud på et socialt sidespor, vurderer lektor ved Danmarks Pædagogiske Universitet Carsten Jessen. Han er midt i et fireårigt forskningsprojekt. Læs artikel

JP-Xtra ◆ Den mobile generation ◆ Mobiltelefon, ja tak ◆ Mobiltelefon, nej tak

JP Afstemning
Er der behov for lovgivning om piercing af børn under 18 år?
Ja
Nej
Ved ikke
Stem

SENESTE NYHED
Livstidsdom til Susan-Morder stadfæstet i Højesteret (14:46)
Telia vil overtage Tele Danmarks mobilvirksomhed (14:37)
EU-Kommissionen: Frankrig skal sikre fri transport (14:29)
Vismandsrapport om Østrig afleveres fredag (14:20)
Lundbeck og Novo forgyldes af dollar-stigning (13:52)

Jyllands-Posten
7 September 2000

1. Web history, web archives — a personal account

1998: Legal deposit of static web (pdf etc.)

2000: The Centre for Internet Studies (CFI)
›promote research on the social and cultural functions and meanings of the internet
›work for the establishing of a national web archive

2001: international conference 'Preserving the Present for the Future — Strategies for the Internet'

1. Web history, web archives — a personal account

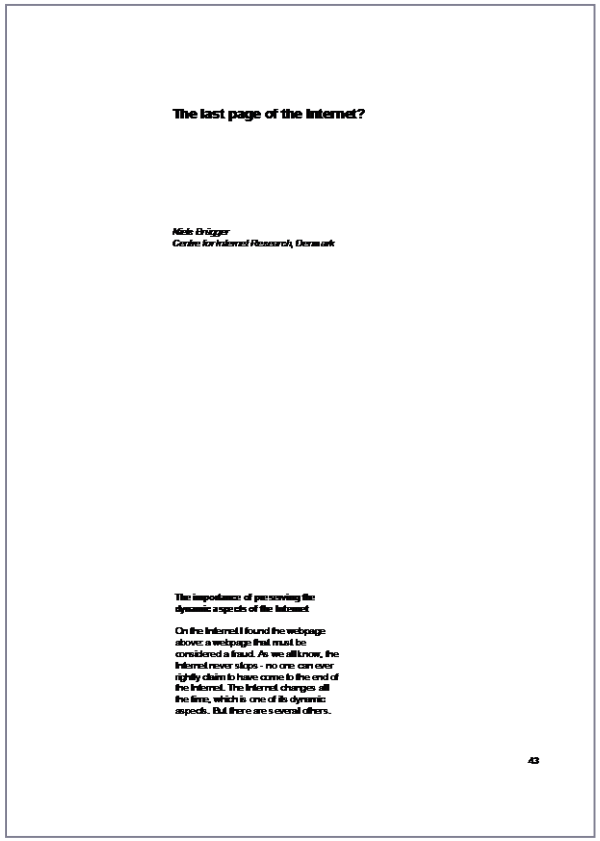
2001-02: netarkivet.dk — a pilot project, the two national libraries and CFI — three archiving strategies

2002-04: Preparation of the revision of the 1998-legal deposit law (the Royal Library & CFI involved)

2005: The Danish web archive Netarkivet established, joint venture between the two national libraries

1. Web history, web archives — a personal account

From vague ideas to more unfolded methodological and theoretical reflections



The last page of the Internet?

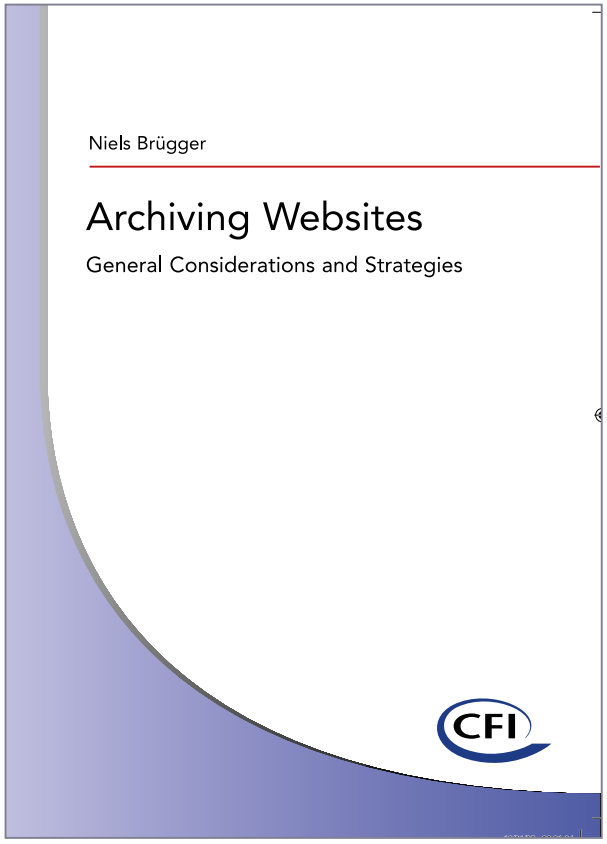
Niels Brügger
Centre for Internet Research, Denmark

The importance of preserving the dynamic aspects of the Internet

On the Internet I found the webpage above: a webpage that must be considered a fraud. As we all know, the Internet never stops - no one can ever rightly claim to have come to the end of the Internet. The Internet changes all the time, which is one of its dynamic aspects. But there are several others.

43

Conference 'Preserving the Present for the Future — Strategies for the Internet', 2001



Niels Brügger

Archiving Websites

General Considerations and Strategies

CFI

2005

... and later several books and articles

1. Web history, web archives — a personal account

Ready to start writing the history of dr.dk 1995-2005 — started in 2007 — still ongoing

But the research infrastructure was still missing — came in 2012

2. Web history — Digital history

Digital history *vs.* Web history:

›Digital history: begin 1990s — web as a tool for making digitized sources available, for dissemination, and for interaction

›Web history: web as a historical source in its own right

2. Web history — Digital history

Web history is an equivocal term

›History with the Web — the Web used as historical source

›History of the Web — historical developments of the Web

3. GENERAL CHARACTERISTICS OF ARCHIVED WEB

Digitized

Previously analog material that has been digitized

Born digital

Has never existed in any other form than digital

Reborn digital

Born digital materiel that has been collected and preserved, and that to a certain degree has been changed in this process

3. GENERAL CHARACTERISTICS OF ARCHIVED WEB

Digital collections
› analog original
› selection of *what* to digitize
› digitization is transparent
› identical copies

Web archives
›ephemeral original
›selection of *what* to archive and of *how* to archive
›archiving less transparent — technical deficiencies and dynamics of updating
›versions without original

3. GENERAL CHARACTERISTICS OF ARCHIVED WEB

Digital collections

› one copy of each
› systematic, register
› hyperlink is add-on

Web archives

› too little, too much
› unsystematic, no register
› hyperlink is inherent

4. THE SCHOLARLY USE OF WEB ARCHIVES

Too little, too much

General inconsistency — inconsistent in terms of time and space

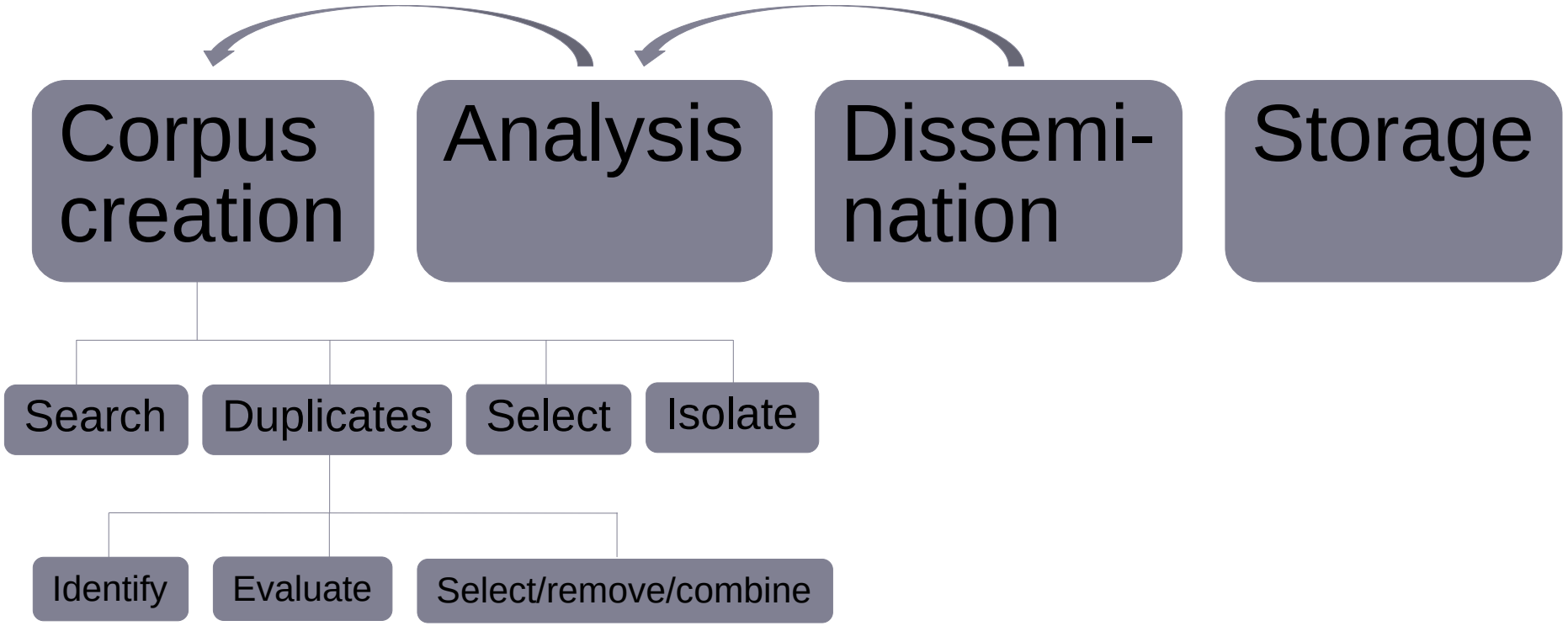
Accumulated complexity and heterogeneity

5. SCHOLARLY NEEDS? A MATRIX FOR A GENERIC APPROACH

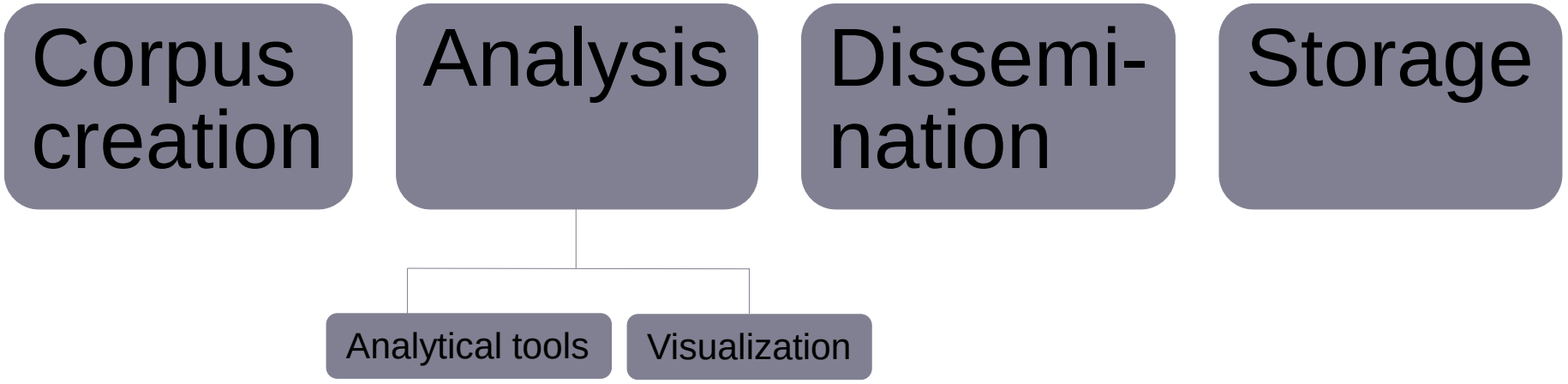
Which tools, procedures, and policies do scholars need?

› The phases of research
› Transversal preconditions

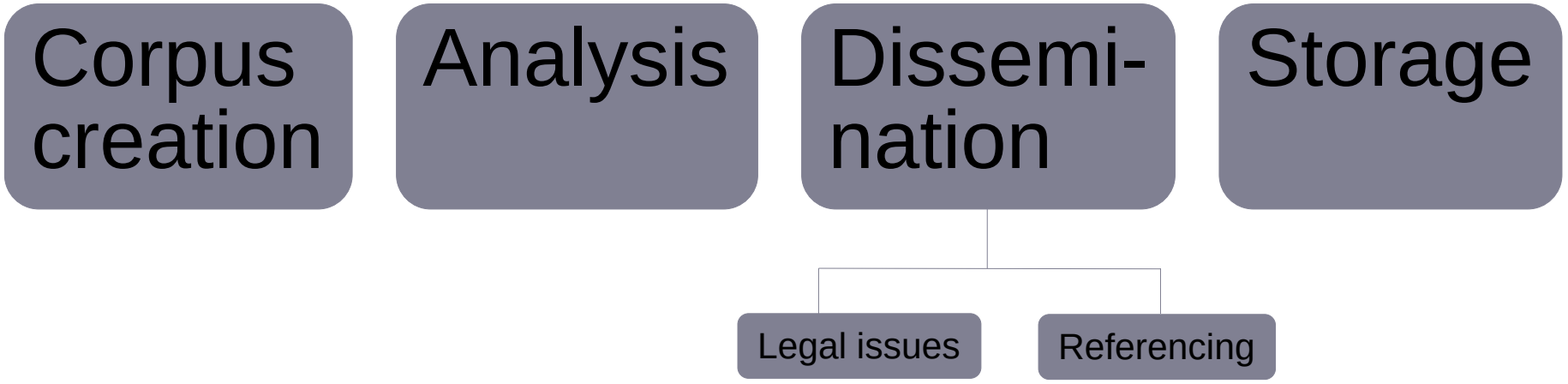
THE PHASES OF RESEARCH



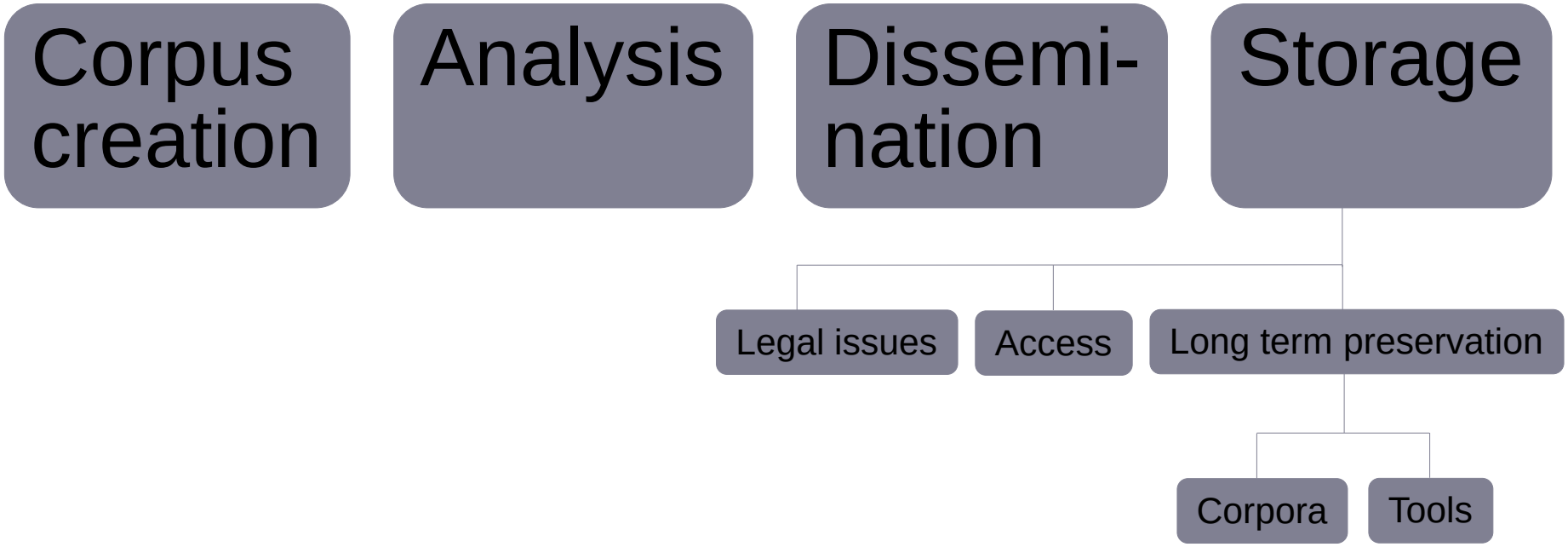
THE PHASES OF RESEARCH



THE PHASES OF RESEARCH



THE PHASES OF RESEARCH



TRANSVERSAL PRECONDITIONS

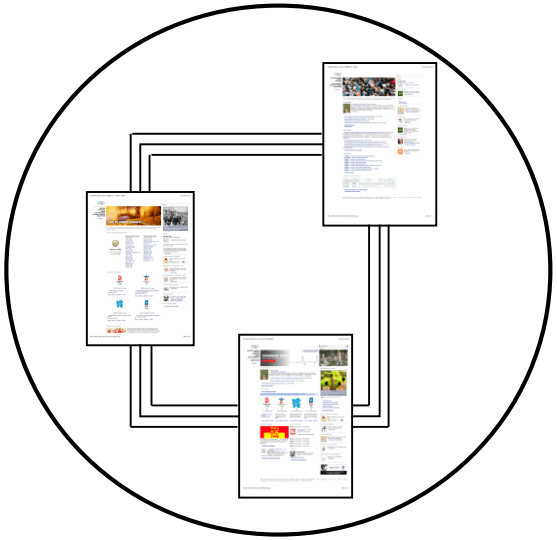
WORKSPACE

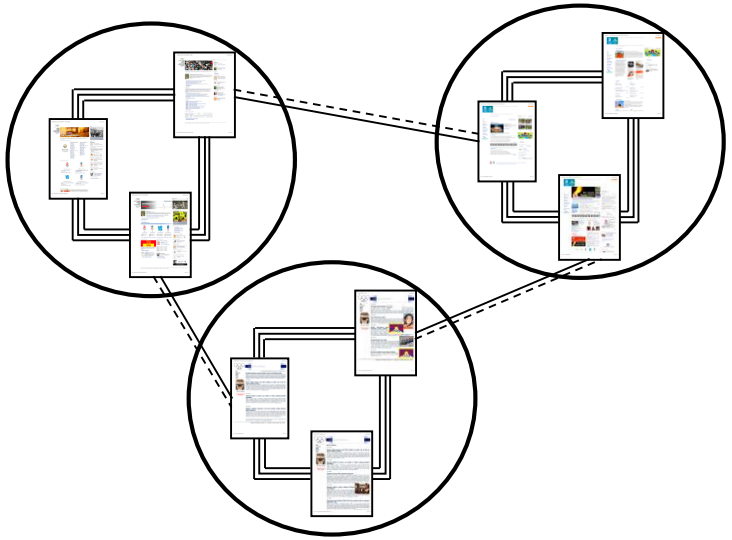
DOCUMENTATION

(RESEARCH) DATA MANAGEMENT

COLLABORATION

6. THE ARCHIVED WEB USED AS HISTORICAL SOURCE





Small data
Close reading

Big data
Distant reading

AARHUS
UNIVERSITY

7. THE HISTORY OF DR.DK 1995-2005


The history of dr.dk 1995-2005

Started in 2007 — still ongoing

7. THE HISTORY OF DR.DK 1995-2005





- institutional analysis of the broadcaster
- technology, the web as a media platform
- content (visible, code)
- structure (menus, website internal hyperlinks)

8. THE DANISH WEB 2005-2015

The main questions:

- What has the entire Danish web looked like in the past, and how has it developed?
- What are the methodological challenges in conducting such a study?
- What kind of research infrastructure do we need to conduct such a study?

8. THE DANISH WEB 2005-2015

Why such a study?

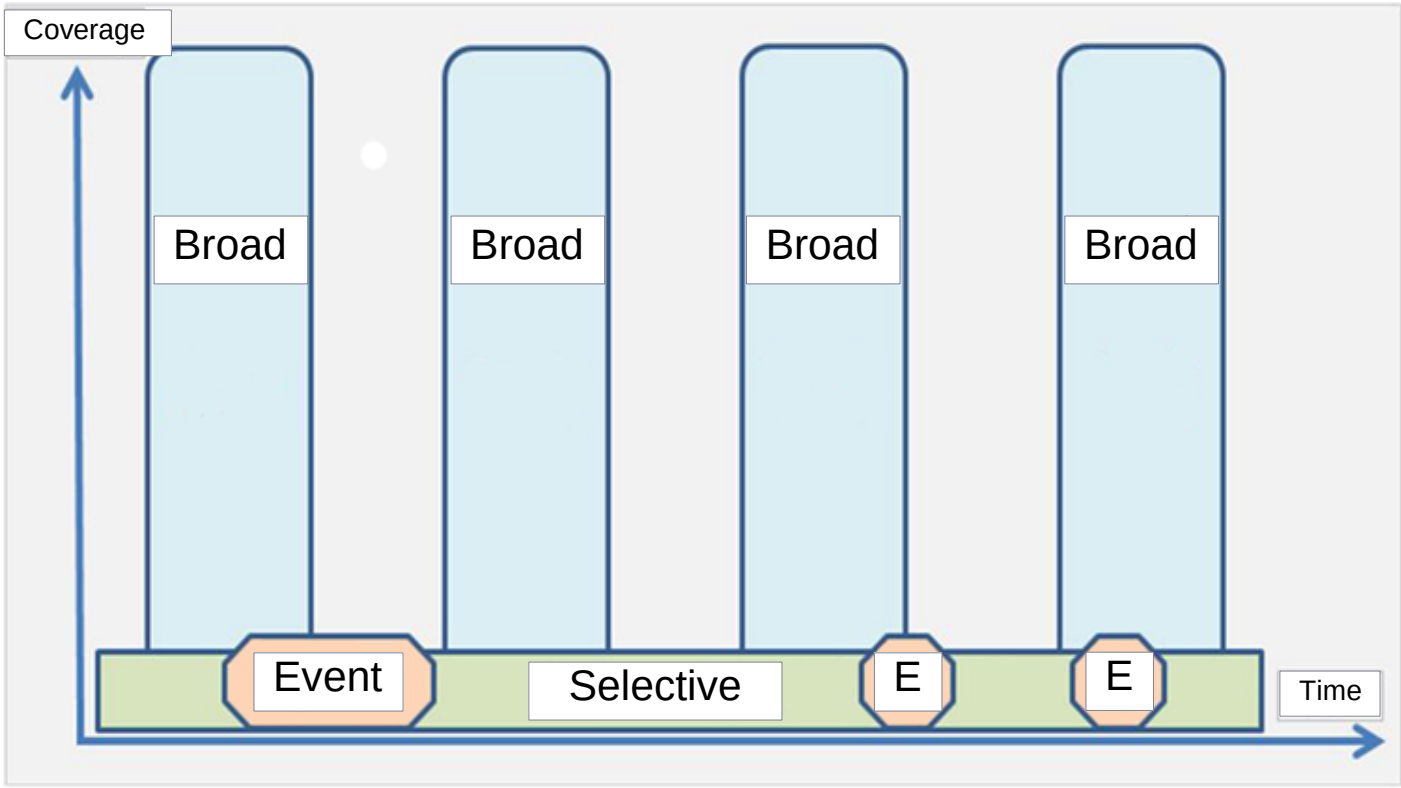
- Back cloth for all other types of Web entities and activities within the national Web area
- Identify some of the patterns of the developments of the Web
- Compare web developments with phenomena outside the Web

8. THE DANISH WEB 2005-2015

What are we studying?

›The Web of the past is gone
›Material in the national Danish Web archive Netarkivet
›Based on Legal Deposit Law, July 2005-
›Web material within the ccTLD .dk and websites on other domains aimed at a Danish audience
›2015: 1.2 million registered domain names within the ccTLD .dk — +600TB

8. THE DANISH WEB 2005-2015



Strategies:

› Broad/bulk
› Selective
› Special collection/event

From http://netarkivet.dk/om-netarkivet

8. THE DANISH WEB 2005-2015

What are we looking for?

› Size
› Space
› Structure
› Aliveness
› Content

8. THE DANISH WEB 2005-2015

Size

- The size of the Danish Web domain (bytes)
- The size of different file types and of file types in general
- The size of websites

8. THE DANISH WEB 2005-2015

Space

› Geolocation of websites
› Searching the text for geographic references, e.g. postcodes in footers

8. THE DANISH WEB 2005-2015

Structure

- Website internal/external hyperlinks
  - Closedness/openness towards the Web
  - Flat/deep websites

- Web domain internal/external hyperlinks
  - Centrality based on in-links
  - Linking of Danish Web domain to the rest of the Web
  - Which other domain names are the most linked-to?

8. THE DANISH WEB 2005-2015

Aliveness

- Domain names: number of new/inactive/disappeared domain names
- Updating: number of Web objects having been changed since last archiving

8. THE DANISH WEB 2005-2015

Content

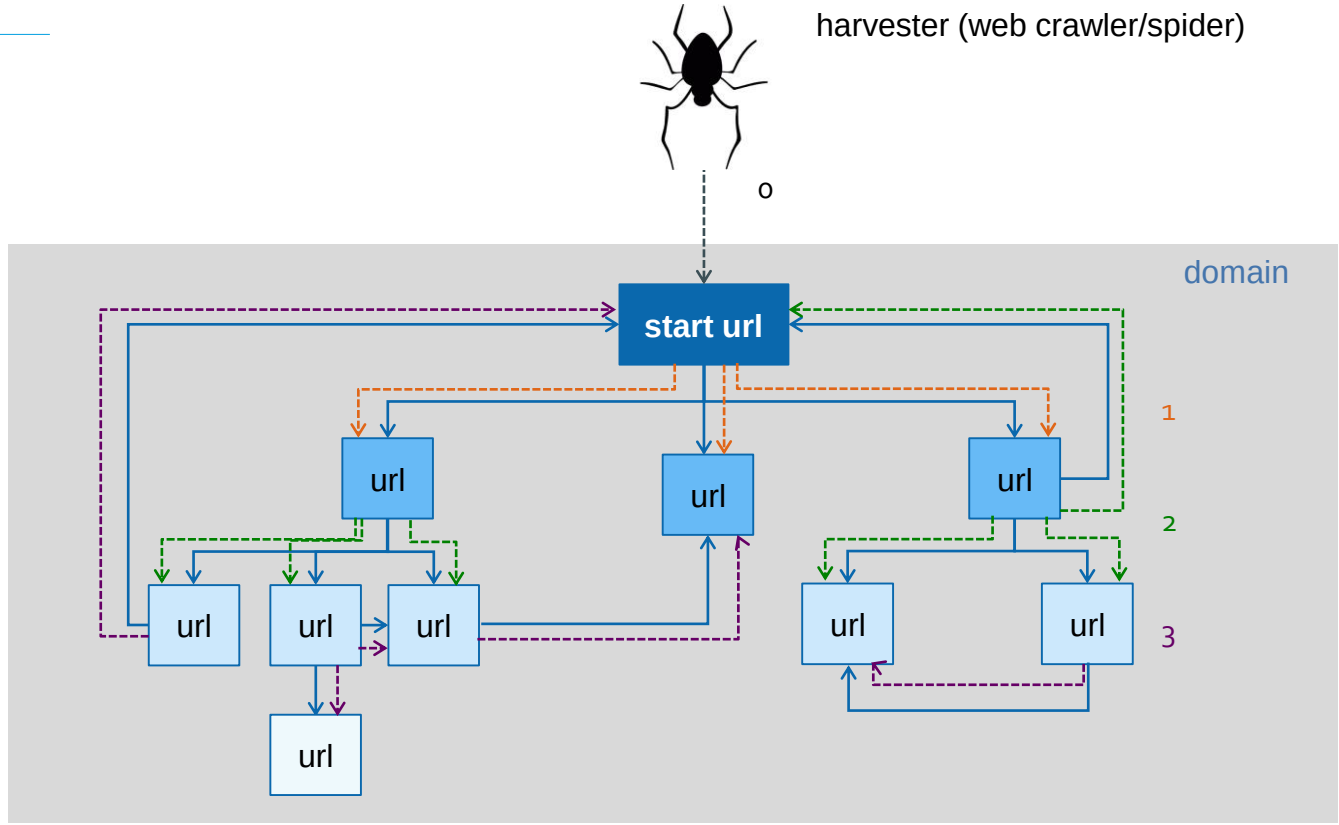
- Number of password protected websites
- Most prevalent file and software types
- The use of Danish, English, German, etc.
- Textual elements on webpages (background color, fonts, length of webpages, placing of menu items, etc.)
- Semantics (word frequencies, where specific issues or topics are to be found on the Danish Web)

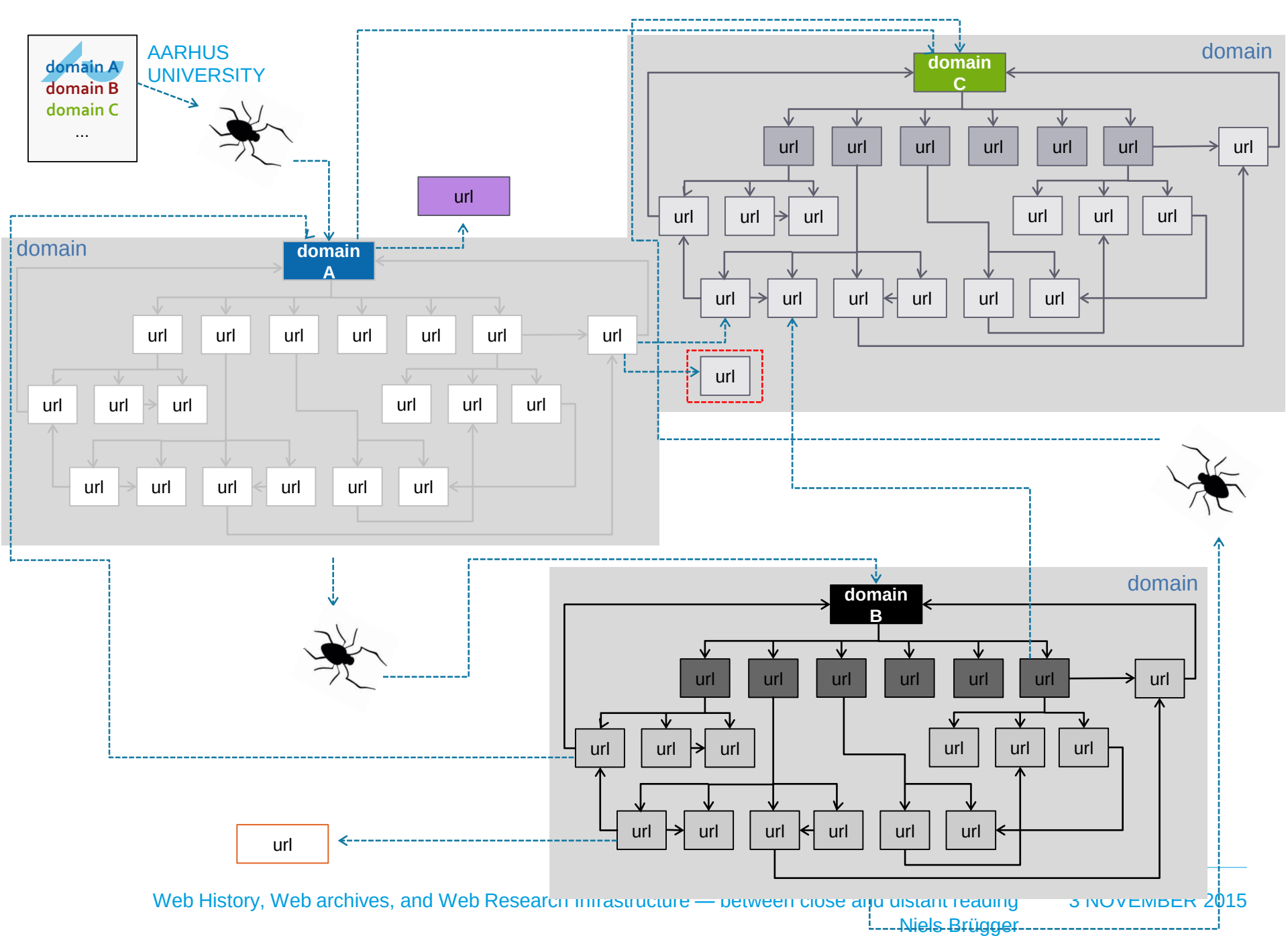
9. ANALYSIS OF BIG ARCHIVED WEB CORPORA: CHALLENGES?

› No 1:1 relation between Danish national archive and the Danish national web domain

› Not everything has been archived
› Unsystematic, no register, no original to compare with
› Archiving takes time, the link structure becomes inconsistent
› Deduplication may affect the subsequent use of the archived material
› Archiving strategies may be changed between two archivings
› Parts of domains may be harvested more than once

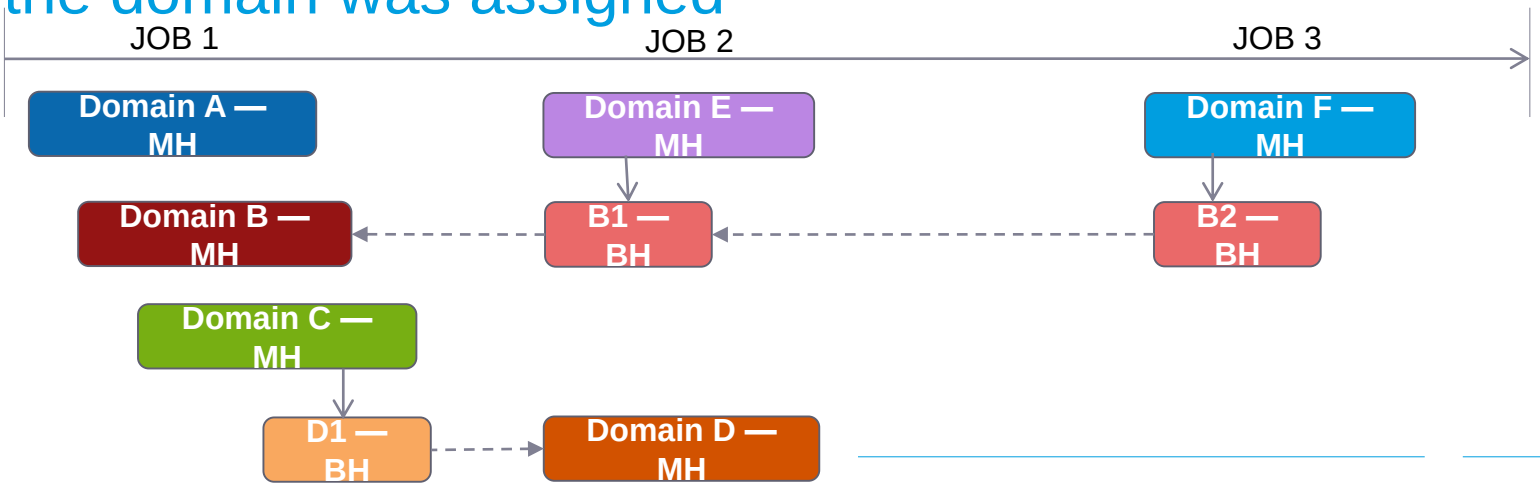
PARTS OF DOMAINS MAY BE HARVESTED MORE THAN ONCE




domain A
domain B
domain C
...
AARHUS UNIVERSITY
domain
domain C
domain A
domain B
url

9. ANALYSIS OF BIG ARCHIVED WEB CORPORA: CHALLENGES?

› Main harvest: objects within a domain which were harvested in the job to which the harvest of the domain was assigned
› By-harvest: objects within a domain which have been harvested in another job than the one to which the harvest of the domain was assigned



9. ANALYSIS OF BIG ARCHIVED WEB CORPORA: CHALLENGES?

Possible solutions

1. Not to use the archive after all
   - › Use the registry of .dk domains

2. Corpus creation
   - › Selection of harvests
   - › Selection of one version of each domain (consisting of the main harvest and possibly by-harvests)

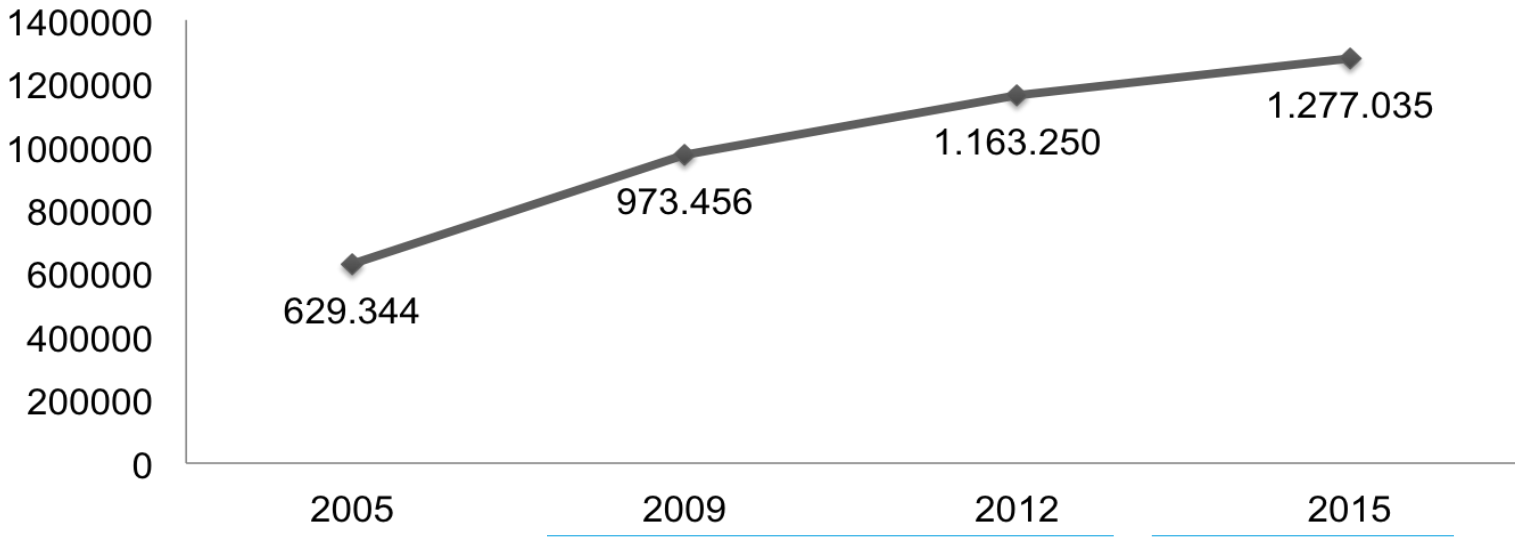
9. ANALYSIS OF BIG ARCHIVED WEB CORPORA: CHALLENGES?

First solution: Registry of .dk domains — aliveness

› What are the total number of domain names over time?
› How many domain names have disappeared compared to the previous years? (and which ones)
› How many domain names have been created compared to the previous year? (and which ones)
› How many domain names have changed hands compared to the previous years? (and which ones)
› How is the relationship of ownership and domains over time?

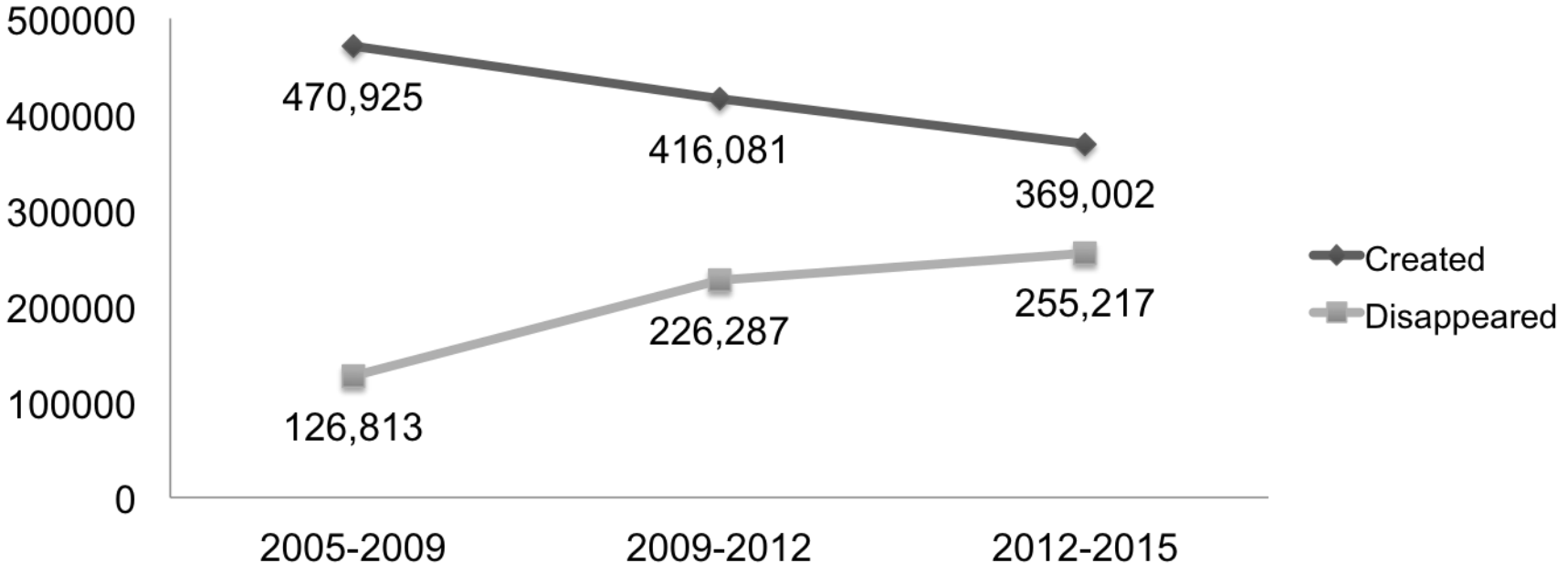
9. ANALYSIS OF BIG ARCHIVED WEB CORPORA: CHALLENGES?

Number of domain names over time (conducted by Netarkivet, thanks to Ditte Laursen, Per Møldrup)



9. ANALYSIS OF BIG ARCHIVED WEB CORPORA: CHALLENGES?

New and disappearing domain names from 2005 to 2015



AARHUS
UNIVERSITY

Year	Domains	Owners	Anonymous
2012	1.163.250	513.326	46.727
2015	1.277.035	549.978	58.710

9. ANALYSIS OF BIG ARCHIVED WEB CORPORA: CHALLENGES?

Number of domain names which have changed hands over time

- In 2015, 14% of the domains from 2012 had changed the owner name
- In 2012 and 2015, less than 10% of the total number of owners owned 50% of the Danish domains
- An observation: If you own more than three domains you are part of the top 10% of domain owners

9. ANALYSIS OF BIG ARCHIVED WEB CORPORA: CHALLENGES?

Second solution: Corpus creation

Collaboration between researchers, curators, IT-developers/-architects and management at the archive

9. ANALYSIS OF BIG ARCHIVED WEB CORPORA: CHALLENGES?

- How is a broad crawl performed? ie. several ”steps”
- When were broad crawls performed?
- How to find the most complete version of a domain within a certain timespan within a broad crawl?
- What do we mean when we talk about a ”web element”, a ”web page”, a ”version” etc.?
- What could a corpus creation algorithm look like?
- How many resources are needed to test and implement a creation of a corpus?

9. ANALYSIS OF BIG ARCHIVED WEB CORPORA: CHALLENGES?

Use of broad crawls

› Internationally recognized as a suitable web harvesting strategy for national archives
› 2-4 broad crawls each year of all domains from .dk as well as Danish websites published under other extensions
› Comprehensive in nature and consistent over time

9. ANALYSIS OF BIG ARCHIVED WEB CORPORA: CHALLENGES?

Selection of broad crawls

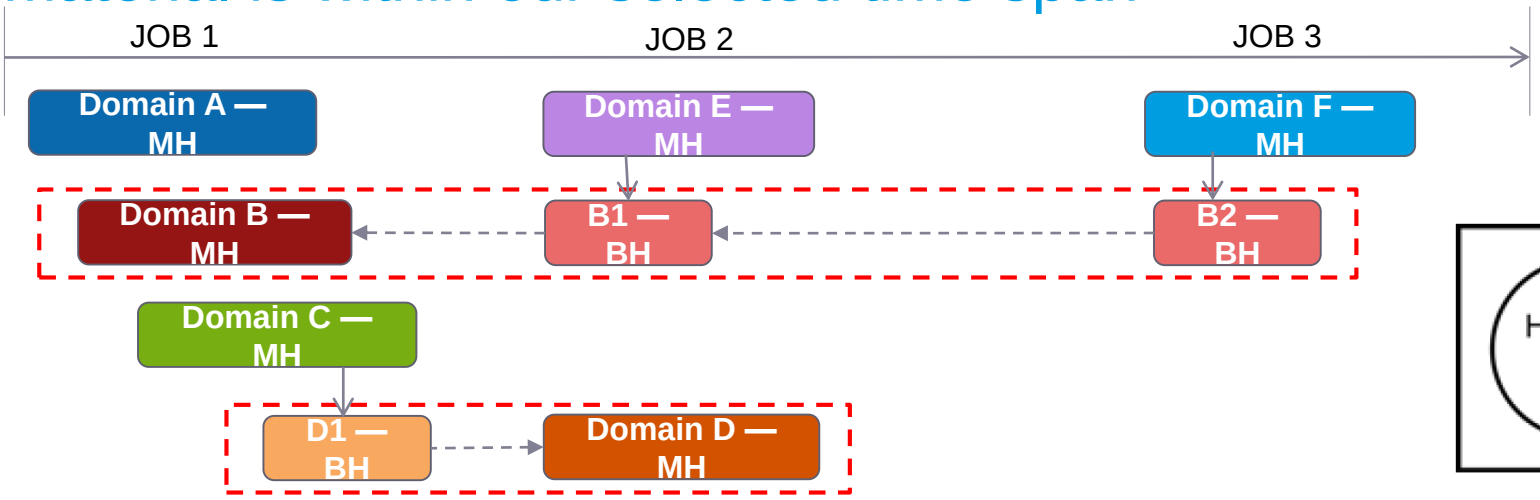
› Four broad crawls, one from each of the years 2006, 2009, 2012 and 2015 (first crawl of the year)

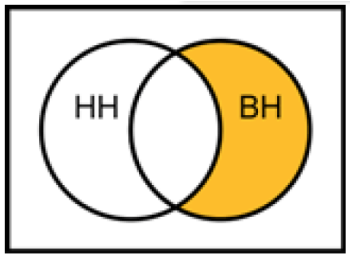
9. ANALYSIS OF BIG ARCHIVED WEB CORPORA: CHALLENGES?

Selection of one havested version of each domain

› Domain version from ’main harvest’

› Inclusion of unique materials from the’ by-harvest’ if the material is within our selected time span

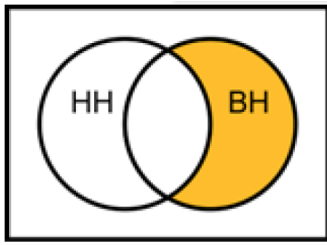




9. ANALYSIS OF BIG ARCHIVED WEB CORPORA: CHALLENGES?

Test of the algorithm

- Tested on the first broad crawl from January 2006 (1TB, only websites <10MB)
- This harvest consists of 127 jobs
- Each job consists of several domains
- We produce a 18GB crawl log enhanced with job IDs



Job ID	Timestamp of the instant of logging	HTTP Status code	Size	URI of the document
2383	2005-12-16T12:57:25.310Z	200	16182	http://www.ottosenfx.dk/page_home/slideshow/slideshow.0002.jpg
2383	2005-12-16T12:57:26.058Z	200	28965	http://www.ottosenfx.dk/page_home/slideshow/slideshow.0010.jpg
2393	2005-12-18T09:53:17.696Z	1	52	dns:www.bartholdy.dk

9. ANALYSIS OF BIG ARCHIVED WEB CORPORA: CHALLENGES?

Test of the algorithm

- Using IBM BigInsights we can perform the algorithm on this large spreadsheet
- The algorithm locates the objects that are not included in a main harvest (’by-harvests’)
- There might be duplicates — in these cases, the algorithm identifies and selects the objects closest to the time of the main harvest

9. ANALYSIS OF BIG ARCHIVED WEB CORPORA: CHALLENGES?

From test to implementation

› How to get from crawl logs to the material that the crawl logs refer to and that we want to analyze? — Should WARC files be opened? Should a subset of an index be used?
› Start making some of the analyzes
› Netarkivet received funding for a 'Cultural Heritage Cluster', hadoop cluster + IBM BigInsights, opened 19 October
› 'The Danish Web 2005-2015' selected as a pilot project

9. ANALYSIS OF BIG ARCHIVED WEB CORPORA: CHALLENGES?

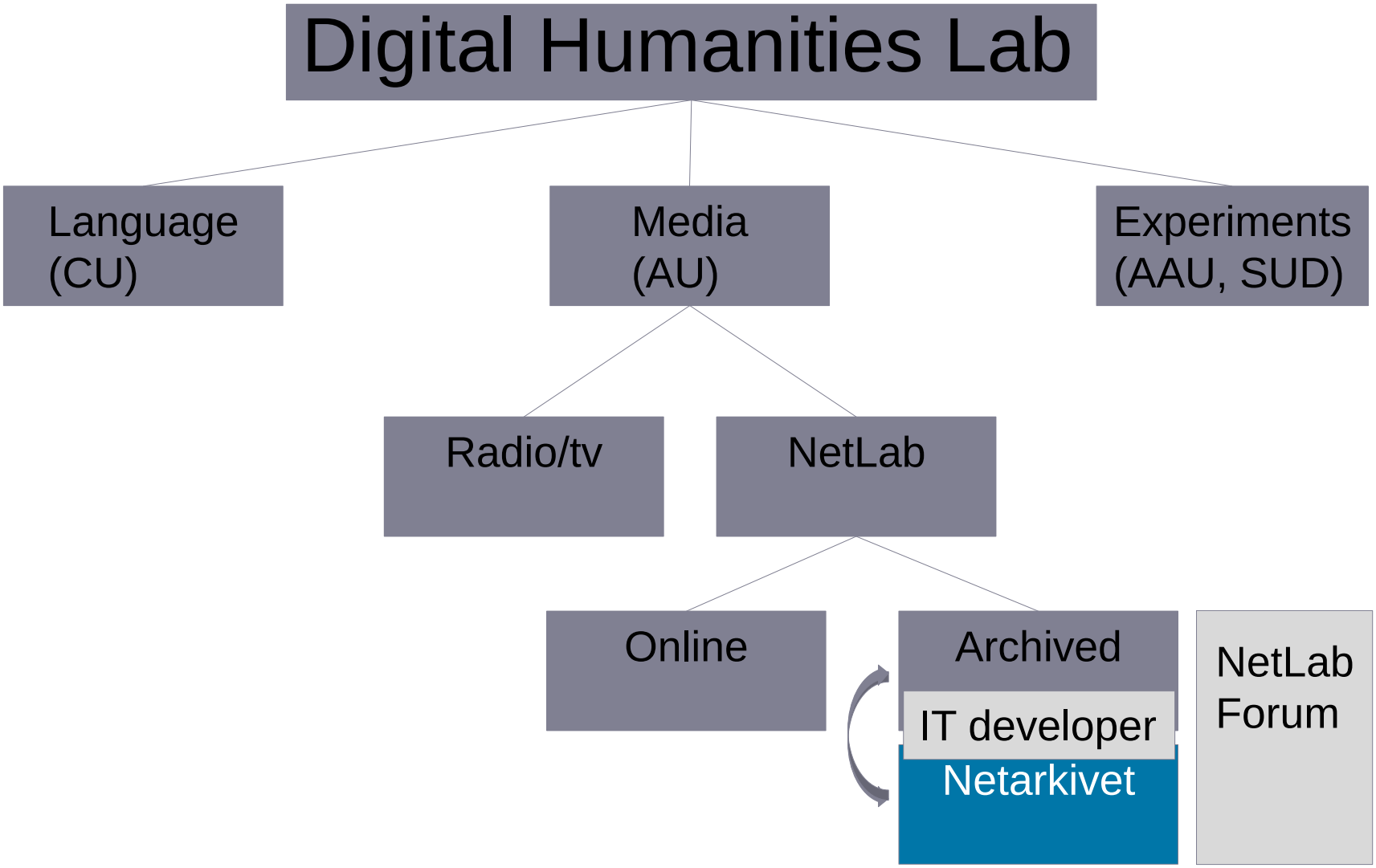
Dissemination and networking

› Book chapters and papers
› An open workshop in Aarhus, Denmark in 2016 for other national web archives and scholars wanting to do similar projects — aiming at establishing transnational ’best practice’ and analytical design

10. WHAT IS NETLAB AND RESAW?

NetLab

- An internet research infrastructure within the Danish research infrastructure for the humanities Digital Humanities Lab
- Established 2012
- Based on the work in the Centre for Internet Studies
- Research(er) driven development of research infrastructure

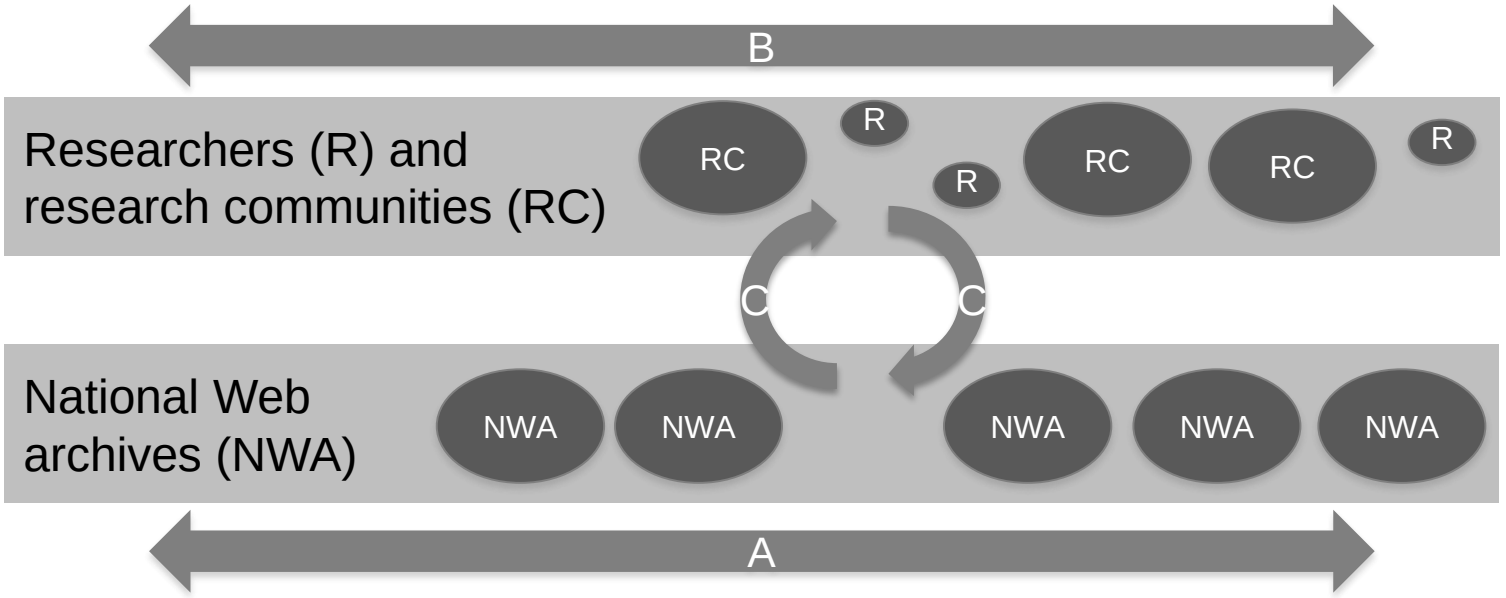

Digital Humanities Lab
Language (CU)
Media (AU)
Experiments (AAU, SUD)
Radio/tv
NetLab
Online
Archived
IT developer
Netarkivet
NetLab Forum

10. WHAT IS NETLAB AND RESAW?

RESAW

›'A Research Infrastructure for the Study of Archived Web Materials' — established in late 2012
›National web archives delimit the borderless information flow on the web by national barriers
›Promote the establishing of a collaborative transnational European research infrastructure for the study of archived web materials

10. WHAT IS NETLAB AND RESAW?



10. WHAT IS NETLAB AND RESAW?

- A consortium of relevant institutions and researchers, European as well as international
- The basis for an application to EU's Horizon 2020 within the topic 'Integrating and opening existing national and regional research infrastructures of European interest'
- resaw.eu

AARHUS
UNIVERSITY

3 NOVEMBER 2015

Web History, Web archives, and Web Research Infrastructure — between close and distant reading

Niels Brügger
PROFESSOR,
HEAD OF THE CENTRE FOR
INTERNET STUDIES, AND OF NETLAB

UNIVERSITET