AARHUS
UNIVERSITY

8 DECEMBER 2016

Analysing national webs — why and what?

NIELS BRÜGGER
PROFESSOR, HEAD OF THE CENTRE
FOR INTERNET STUDIES, AND OF NETLAB

VERSITUNI

AARHUS
UNIVERSITY

AGENDA — STUDYING NATIONAL WEB DOMAINS

1. Why?
2. A brief research history
3. Sources?
4. Challenges

AARHUS
UNIVERSITY

1. WHY?

- **historical knowledge** about the development of the (trans-)national webs — in its own right, as backdrop for web studies, and as element in other studies
- develop adequate **methods** for Big Data cultural heritage studies (if based on archived web)
- develop adequate **analytical tools and research infrastructures** for Big Data cultural heritage studies
- gain **experience** with working with large quantities of archived web — play with the material, methods, RQs

AARHUS
UNIVERSITY

2. A BRIEF RESEARCH HISTORY

2000	First studies of national webs	Technical focus (web & web archives), rarely historical
2010-13	**U of Sydney**: Internet History in Australia and the Asia-Pacific	Internet (not web only), historical, not archived web
2012-14	**OII (BL/JISC)**: Big Data: Demonstrating the Value of the UK Web Domain Dataset for Social Science Research	Web, historical, archived web, link analysis
2013-	**AU/Netarkivet**: Probing a nation's web sphere – the historical development of the Danish web	Web, historical, archived web, broad
2014-15	**UoLondon, OII, BL, AU**: Big UK Domain Data for the Arts and Humanities (BUDDAH)	Web, historical, archived web, broad
2014-15	**UoWaterloo**: A Longitudinal Analysis of the Canadian World Wide Web as a Historical Resource	Web, historical, archived web, broad
2014-17	**ISCC – CNRS, and more**: Web90 – Patrimoine, Mémoires et Histoire du Web dans les années 1990	Web, historical, archived web — and more —, broad
2014-15	**Anat Ben-David**: What does the web remember of its deleted past? An archival reconstruction of the former Yugoslav top-level domain	Web, historical, archived web

N. Brügger (2017, forthc.). Probing a nation's web domain: A new approach to web history and a new kind of historical source. In G. Goggin & M. McLelland (Eds.), *Routledge Companion to Global Internet Histories*. New York: Routledge.

3. SOURCES?

Three main approaches, regarding sources:

› the **archived** web in itself

› **other** types of sources (user statistics, reports, documents, interviews...)

› or combinations

AARHUS
UNIVERSITY

4. CHALLENGES

a. Does it make sense to talk about national webs?
b. Delimiting a nation — on the web
c. Dynamics of ccTLD lists
d. The archived web never covers the online web
e. Creating a corpus, too much — versions?
f. Creating a corpus, too little — small collections?
g. Creating a corpus, nothing — no collections?
h. Creating a corpus, what does it look like?
i. Knowing the web archive, data cleaning/engineering
j. Analytical tools, methods, visualisation

AARHUS
UNIVERSITY

4a. DOES IT MAKE SENSE TO TALK ABOUT NATIONAL WEBS?

› potentially **global** — actually **national/local**
› **users**: survey *The media menus of Danish internet users 2009*, the web is largely used as a national (or even local) medium, the ”national horizon is strong except for ‘Searching’ info about and ‘Buying goods’” (Finnemann et al. 2009, pp. 31-35)
› **content providers**: comparative study of news websites in nine different countries, concluded that online news is strongly nation-centred (Curran et al. 2013, pp. 887-891).

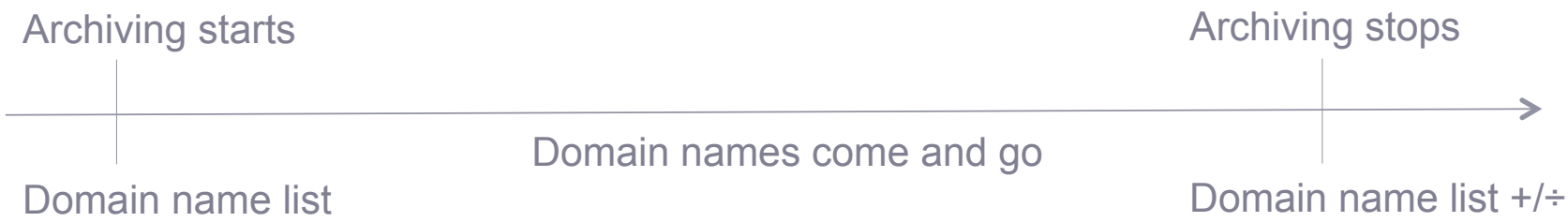
AARHUS
UNIVERSITY

4b. DELIMITING A NATION — ON THE WEB

- **geo-physical** space (analog broadcasting) — **cultural** space
- **web** space — the web's own institutionalized national delimitations, ccTLD:
  - ccTLD — automatic
  - ccTLD+ — automatic & manual
  - ÷ccTLD — automatic & manual
- language, but only in a limited number of countries

4c. THE DYNAMICS OF ccTLD LISTS

› can we use the official **ccTLD list as a register**, when dealing with the archived web?
› **yes**, at the date the list was created
› **no**, because the list is dynamic, and archiving an entire web domain takes time



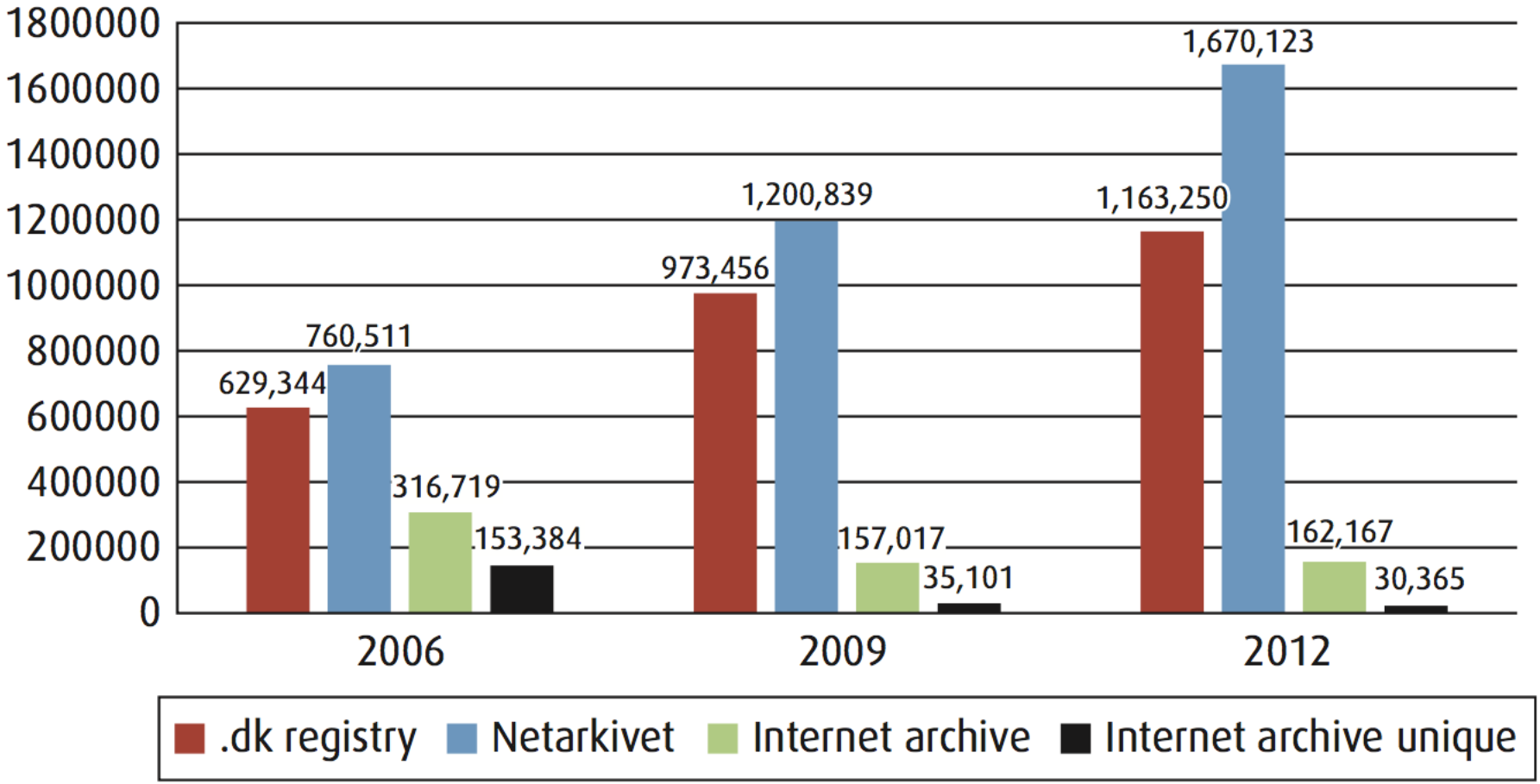


Figure 3.8 Domain names in the Internet Archive not found in the .dk registry

N. Brügger, D. Laursen, J. Nielsen (2017, forthc.). Exploring the domain names of the Danish web. In N. Brügger & R. Schroeder (Eds.), *The Web as History: Using Web Archives to Understand the Past and the Present*. London: UCL Press.

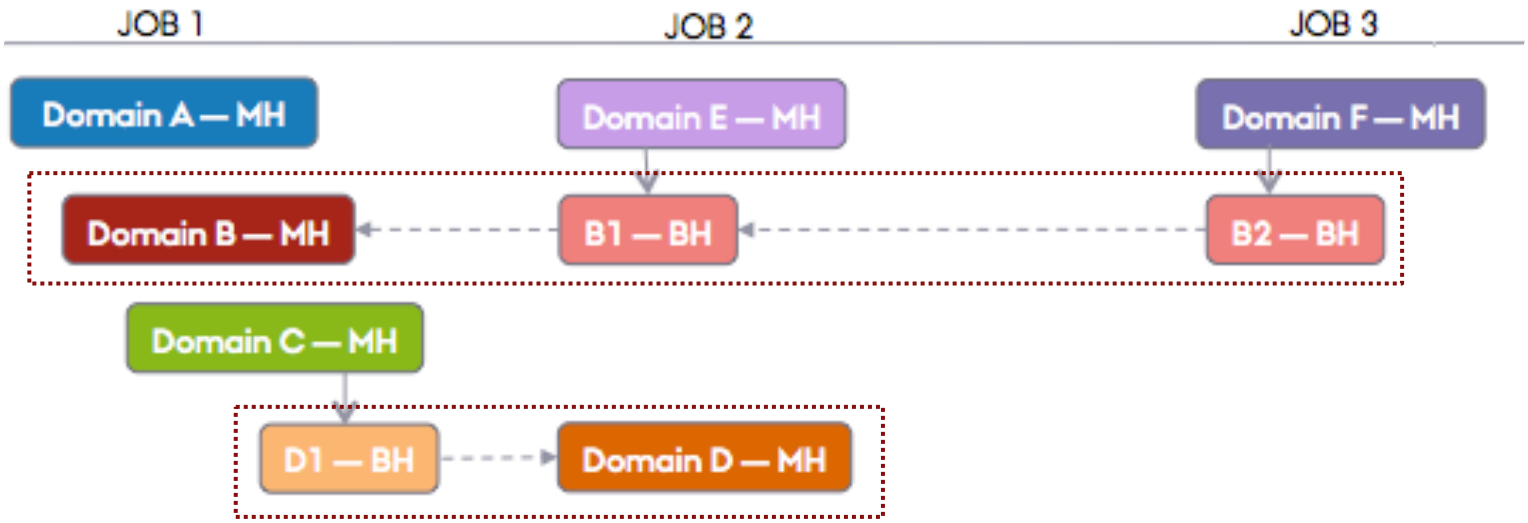
4d. THE ARCHIVED WEB NEVER COVERS THE ONLINE WEB

› the archived web is **never a 1:1 copy** (technical reasons, dynamics of updating)
› we tend to **analyse the web archive** and not the live web as it was

4e. CREATING A CORPUS, TOO MUCH — VERSIONS?

› web crawling by following links crates **more versions** of ‘the same’
› how to find out **if, where, and how much** this is the case?
› how to deal with it?

4e. CREATING A CORPUS, TOO MUCH — VERSIONS?



4f. CREATING A CORPUS, TOO LITTLE— SMALL COLLECTIONS?

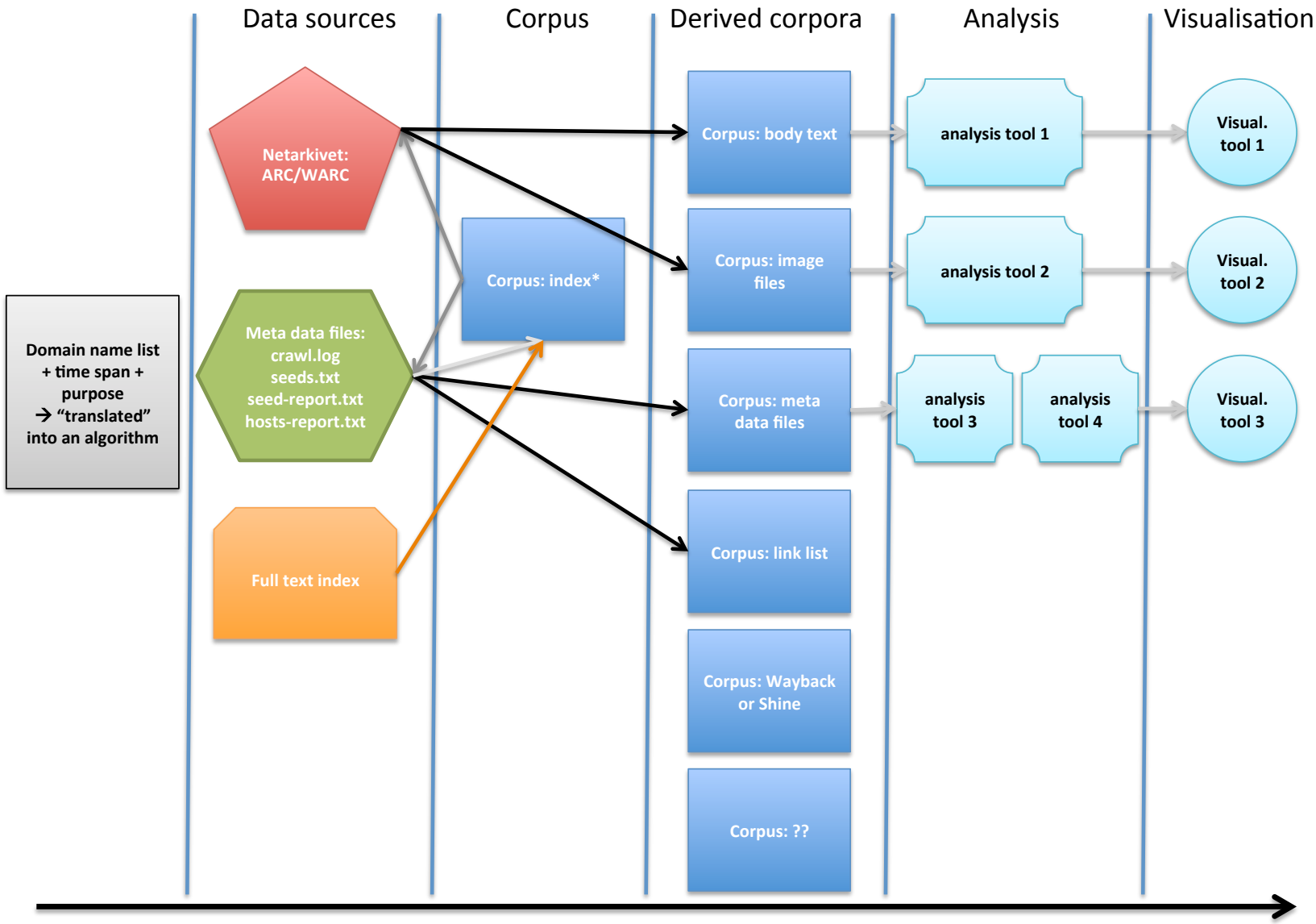
› what do countries with **no broad national crawls** do? (the Netherlands, Australia...)
› use the **ccTLD-list as yardstick** (although it is dynamic)?
› combine own collections with content from **other web archives** — the challenge of combining (different archiving strategies, quality, settings...)?

4g. CREATING A CORPUS, NOTHING — NO COLLECTIONS?

- what do countries with **no national web archive do** (Belgium, Italy, Poland...)?
- or what about **countries that do no longer exist** (Yogoslavia...)?
- use the **ccTLD-list as yardstick** (although it is dynamic)?
- combine content from **different web archives** — the challenge of combining (different archiving strategies, quality, settings...)?

AARHUS
UNIVERSITY

4h. CREATING A CORPUS, WHAT DOES IT LOOK LIKE?


Data sources
Corpus
Derived corpora
Analysis
Visualisation
Domain name list + time span + purpose → "translated" into an algorithm
Netarkivet: ARC/WARC
Meta data files: crawl.log seeds.txt seed-report.txt hosts-report.txt
Full text index
Corpus: index*
Corpus: body text
Corpus: image files
Corpus: meta data files
Corpus: link list
Corpus: Wayback or Shine
Corpus: ??
analysis tool 1
analysis tool 2
analysis tool 3
analysis tool 4
Visual. tool 1
Visual. tool 2
Visual. tool 3

AARHUS
UNIVERSITY

4i. KNOWING THE WEB ARCHIVE — DATA CLEANING/ENGINEERING

› web archives not **prepared for Big Data studies**
› the web is messy, web archives tend to **add their own messiness**, rarely documented in any systematic way
› finding out **what is in the web archive**, how it is structured, known deficiencies, deficiences discovered when working with the material
› Big Data = trends, patterns, not account for all data

› make as informed choices as possible

AARHUS
UNIVERSITY

4j. ANALYTICAL TOOLS, METHODS, VISUALISATION

› which **tools** can be used, and do they scale?
› known **methods** may need to be revised
› new ways of **visualising** may be needed, ‘explorative analytical visualisation’ and ‘end-user visualisation’

4. CHALLENGES — SUMMARY

› **nesting** of challenges
› a **new type** of **source criticism** is needed, acknowledging the specificities of the archived web — collaboration between web archivists and researchers
› find ways of making as **informed choices** as possible

AARHUS
UNIVERSITY

8 DECEMBER 2016

Analysing national webs — why and what?

NIELS BRÜGGER
PROFESSOR, HEAD OF THE CENTRE
FOR INTERNET STUDIES, AND OF NETLAB

VERSITUNI