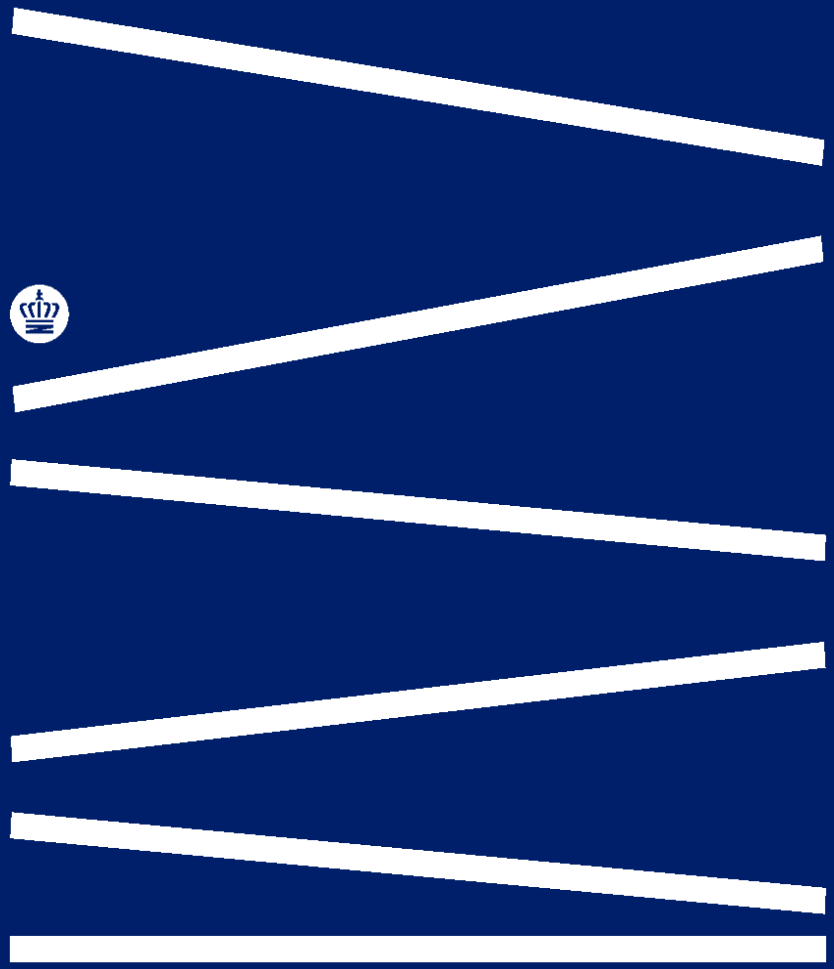
ESTABLISHING A CORPUS OF THE ARCHIVED WEB

The case of the Danish web from 2005 to 2015

Niels Brügger, Aarhus University, DK
Ditte Laursen, Royal Danish Library
Janne Nielsen, Aarhus University, DK
20190606 IIPC, Zagreb, Croatia

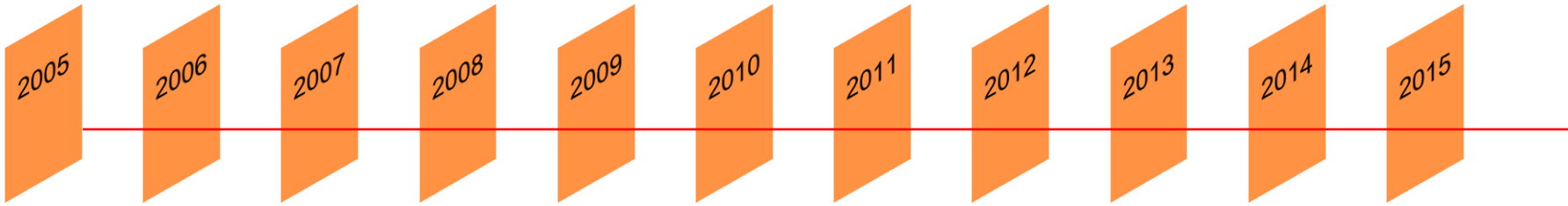


DET KGL. BIBLIOTEK
Royal Danish Library

AARHUS UNIVERSITY

Large scale web history: a history of the Danish web

Aim of "Probing a Nation's Web Domain":
—to analyse the historical development of the entire Danish web domain
—to develop Big Data methods on the archived web
—to establish methods to create a corpus



Brügger, N. (2017). 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.), *The Routledge Companion to Global Internet Histories* (p. 61-73). New York/Abingdon : Routledge.

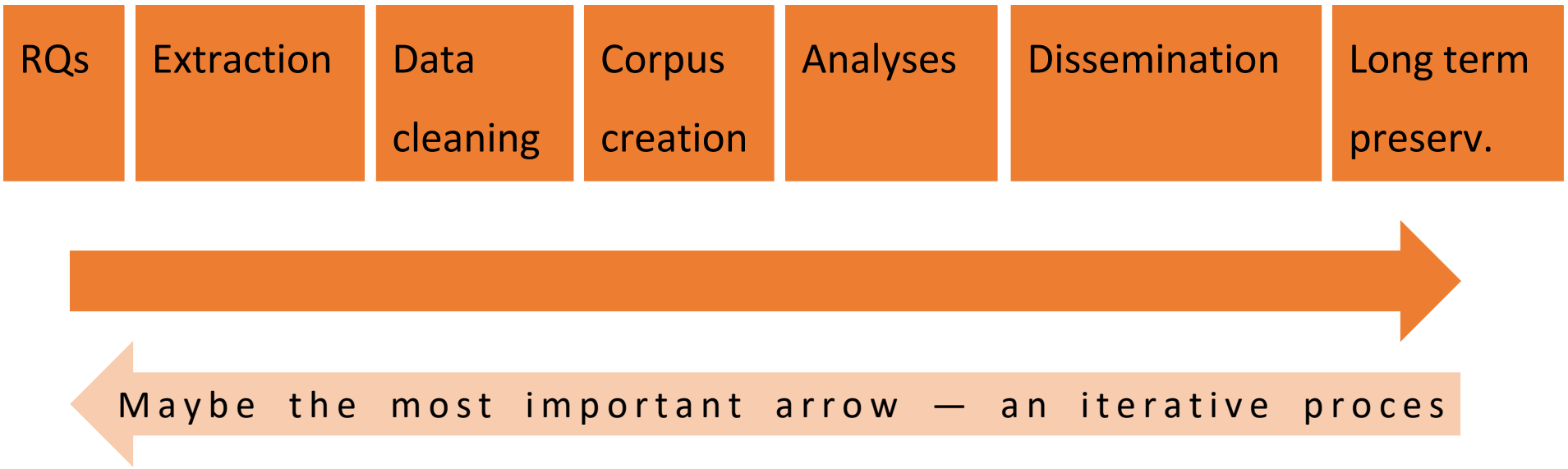
Large scale web history: a history of the Danish web

—Data: material from the national Danish web archive Netarkivet
—Hardware: the DeIC National Cultural Heritage Cluster at the Royal Danish Library, computes 70TB
—The project team: Niels Brügger (Aarhus University, NetLab), Janne Nielsen (Aarhus University, NetLab), Ulrich Have (Aarhus University, NetLab), Per Møldrup-Dalum (Royal Danish Library) and me
—Project start: 2014

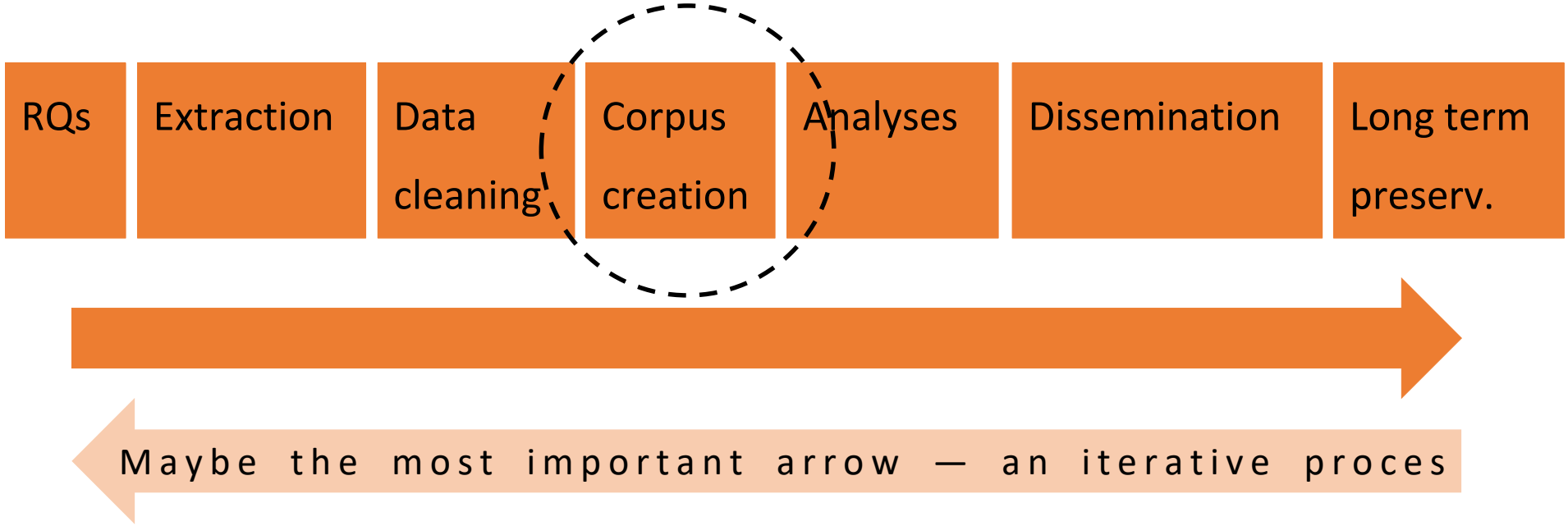
What has the Danish web looked like, and how has it developed?

Initial long list (short version):
—Size
—Space (geolocation)
—Structure (networks of hyperlinks)
—Vivacity (updating)
—Content (ie. closedness, file and software types, language, and semantics)

The overall work phases

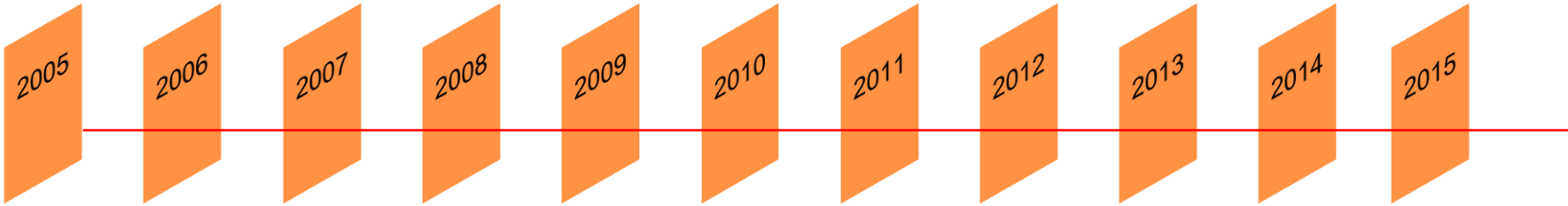


The overall work phases



Corpus creation

— first broad crawl of each year (step 1+2 including special harvest ‘very big sites’, ‘ministeries and departments’)

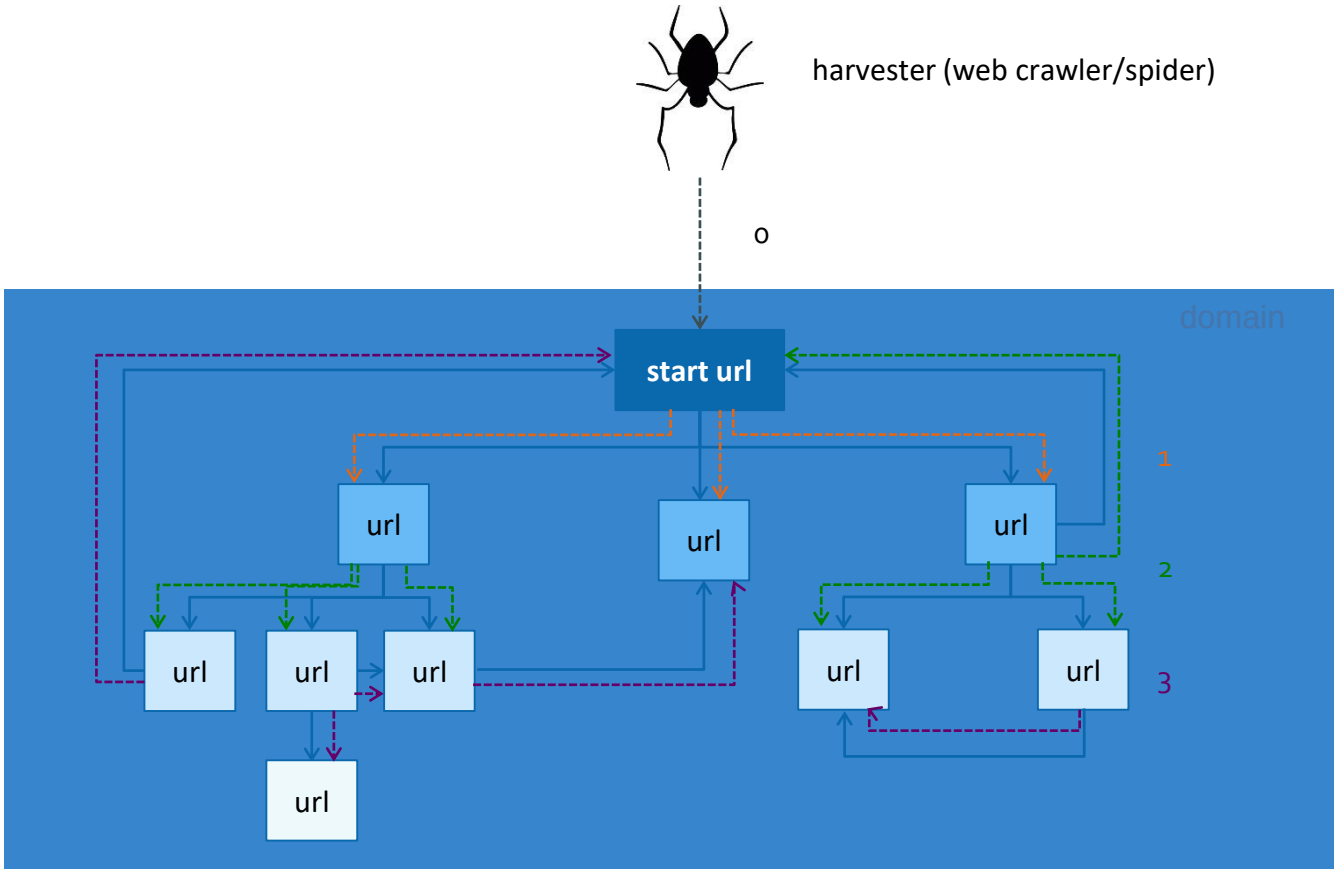


Corpus creation

— reduce that parts of domains may be harvested more than once

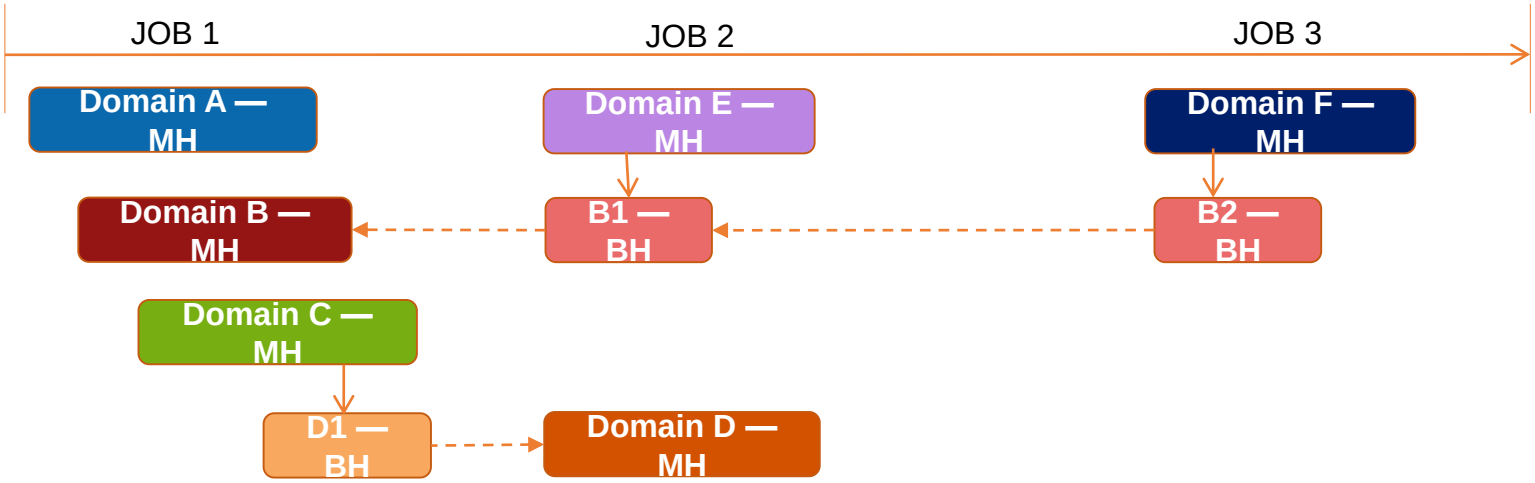
Corpus creation

— reduce that parts of domains may be harvested more than once



Corpus creation

— reduce that parts of domains may be harvested more than once



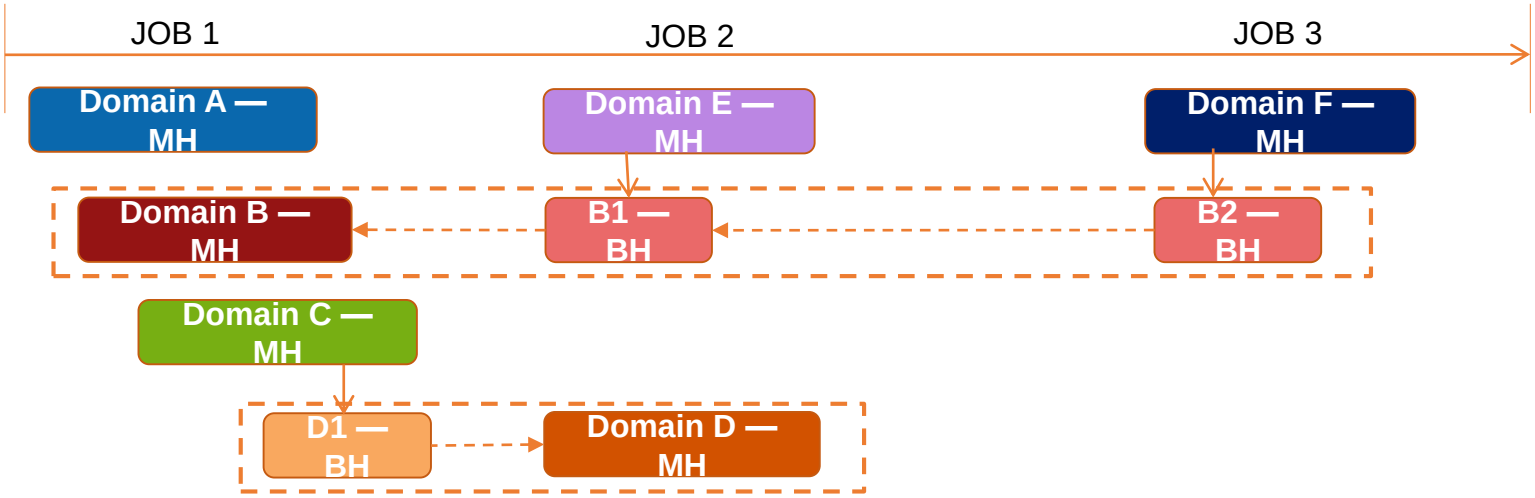
Corpus creation

— reduce that parts of domains may be harvested more than once

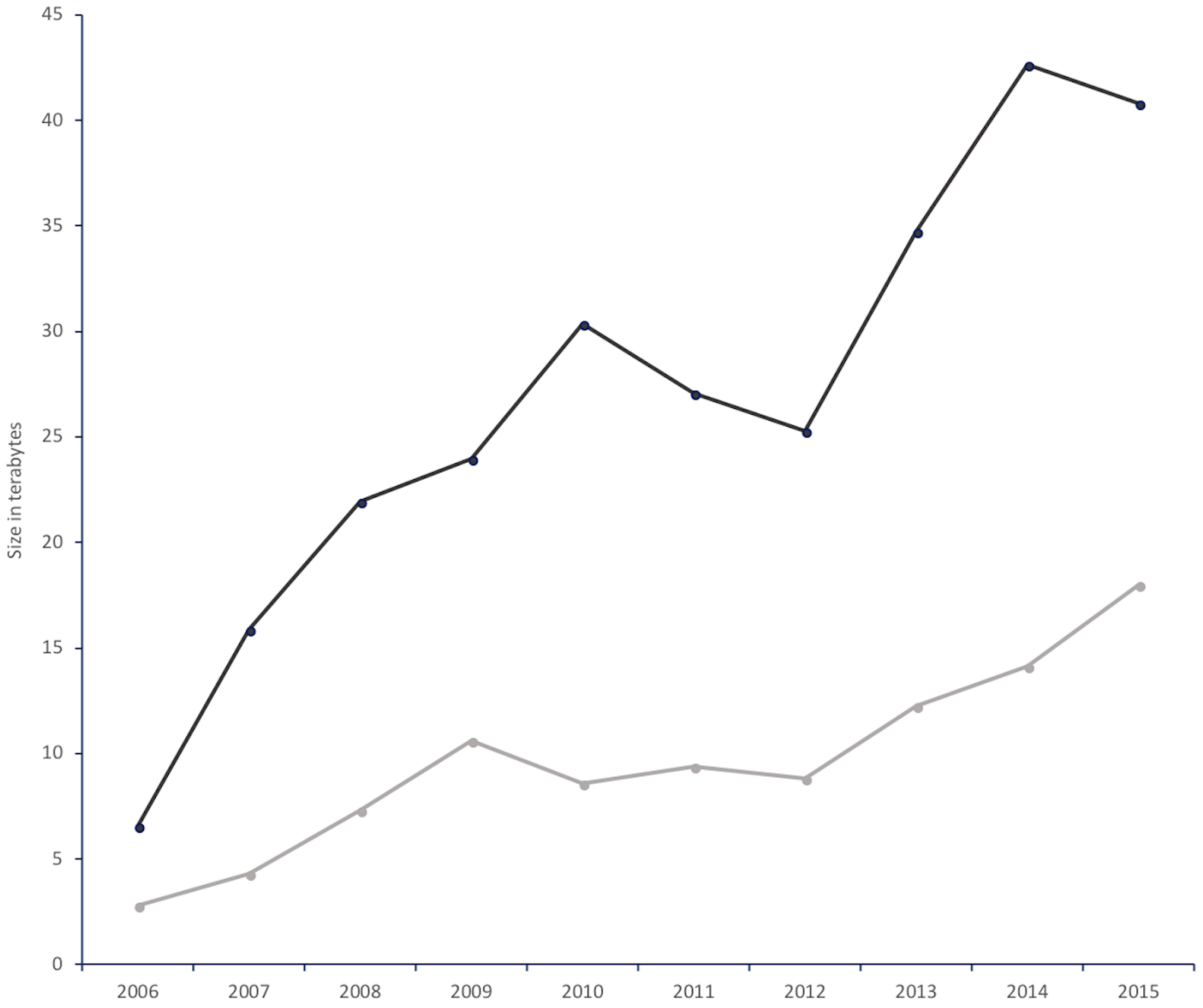
Selection of one havested version of each domain

–Domain version from ’main harvest’

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



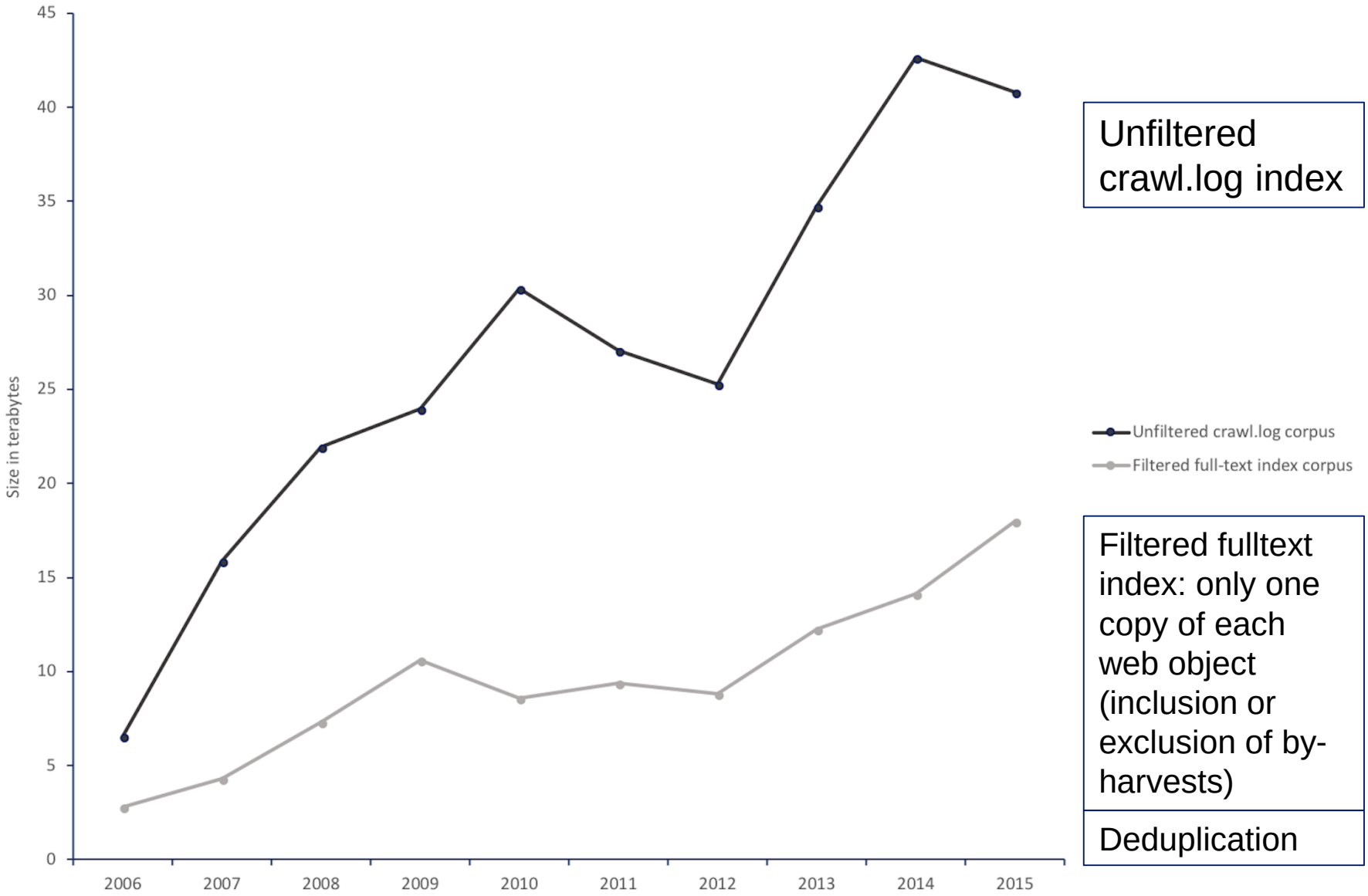
Unfiltered vs. filtered corpus



Unfiltered crawl.log index

Filtered fulltext index: only one copy of each web object (inclusion or exclusion of by-harvests)

Unfiltered vs. filtered corpus



Unfiltered crawl.log index

Filtered fulltext index: only one copy of each web object (inclusion or exclusion of by-harvests)

Deduplication

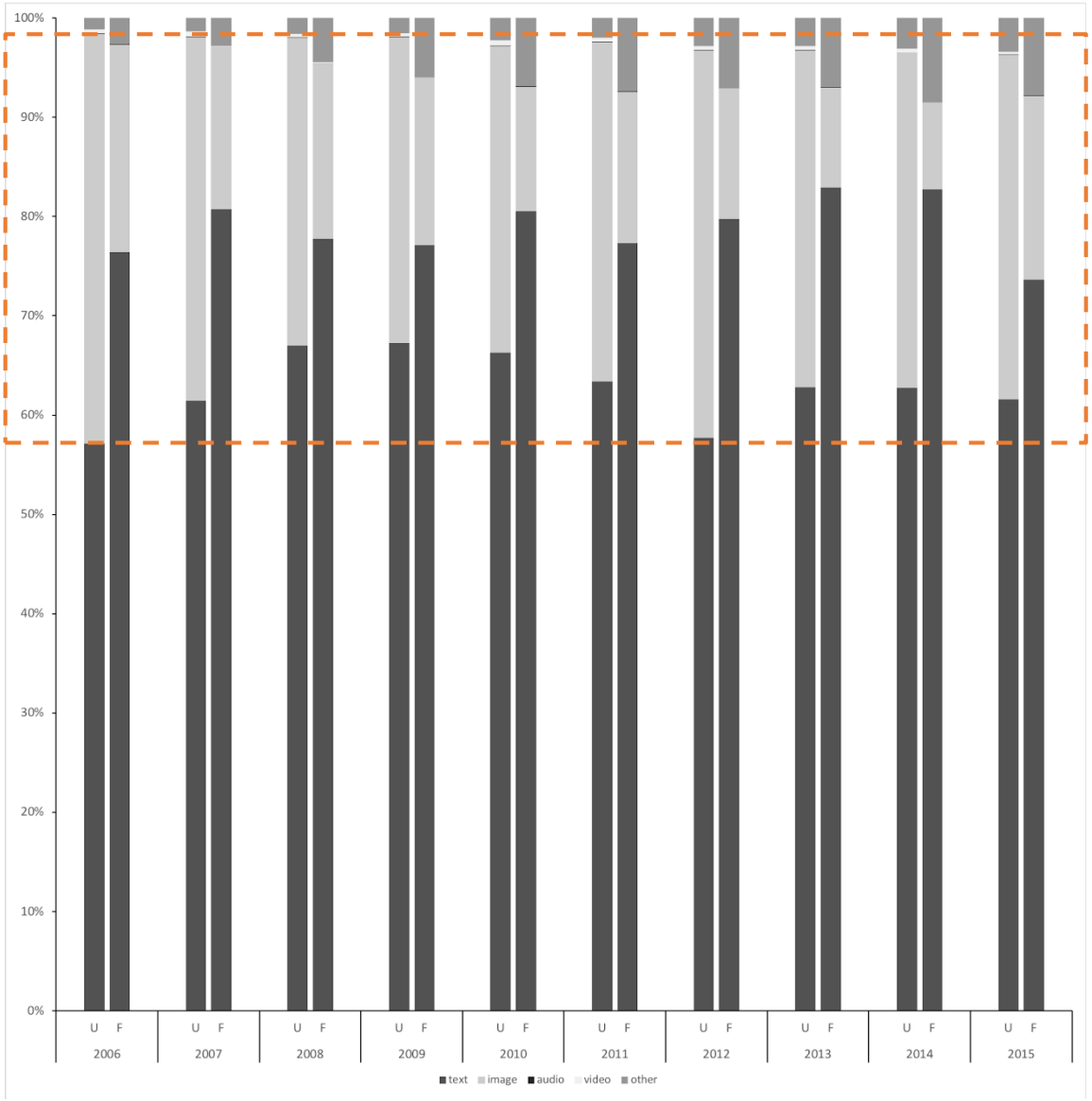
Unfiltered vs. filtered corpus

Unfiltered crawl.log index

Filtered fulltext index: only one copy of each web object (inclusion or exclusion of by-harvests)

Deduplication

Unfiltered vs. filtered corpus

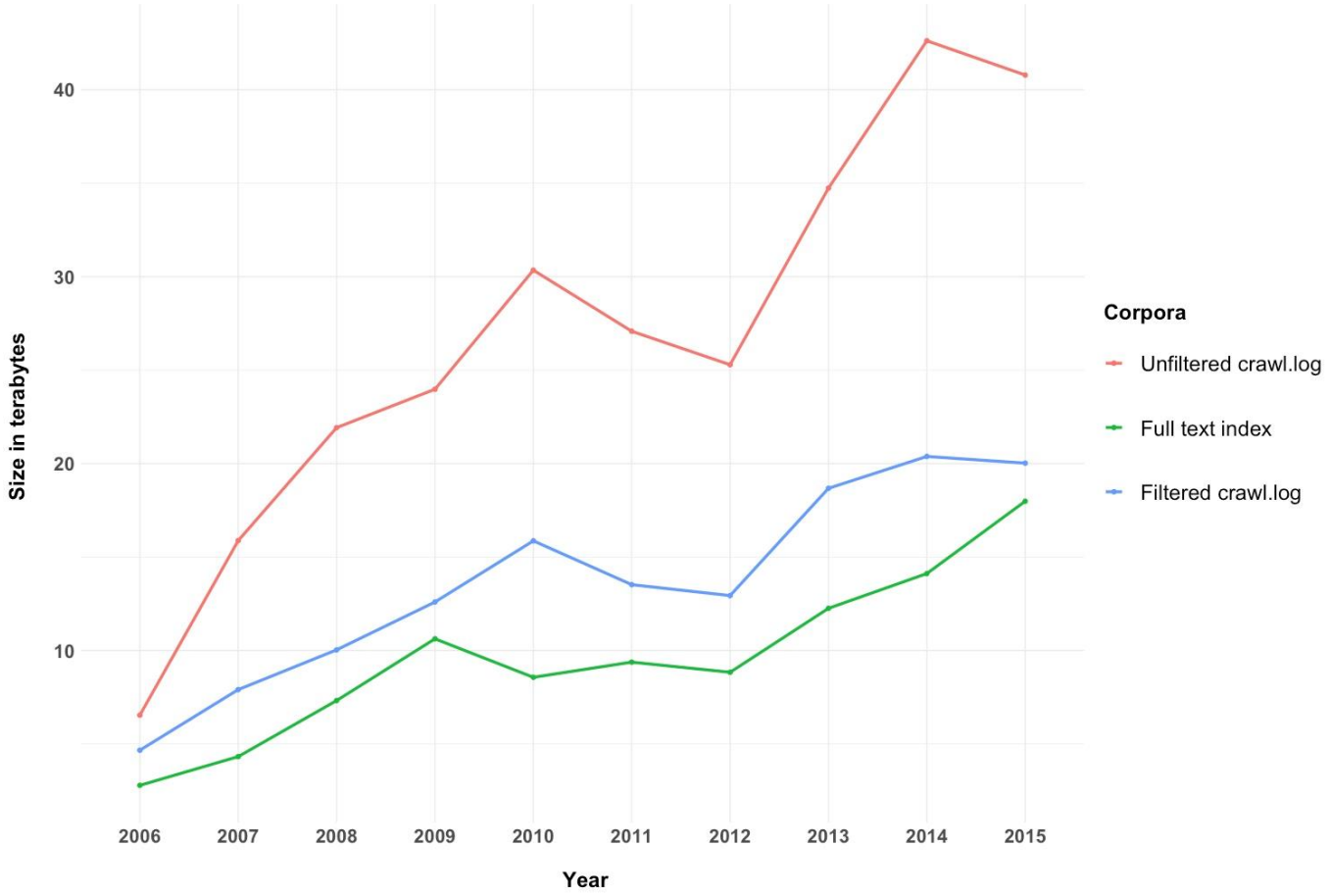


Deduplication

Corpus creation

—Archives are biased, but this can be mitigated by creating a corpus
—A filtered corpus should be able to generate the most valid results, compared to an unfiltered corpus
—The method for creating the corpus is very important and greatly influences the results
—We do not know the importance of different factors and in what way they might influence depending on different analytical and technological approaches

Corpus creation 2.0 pre-view



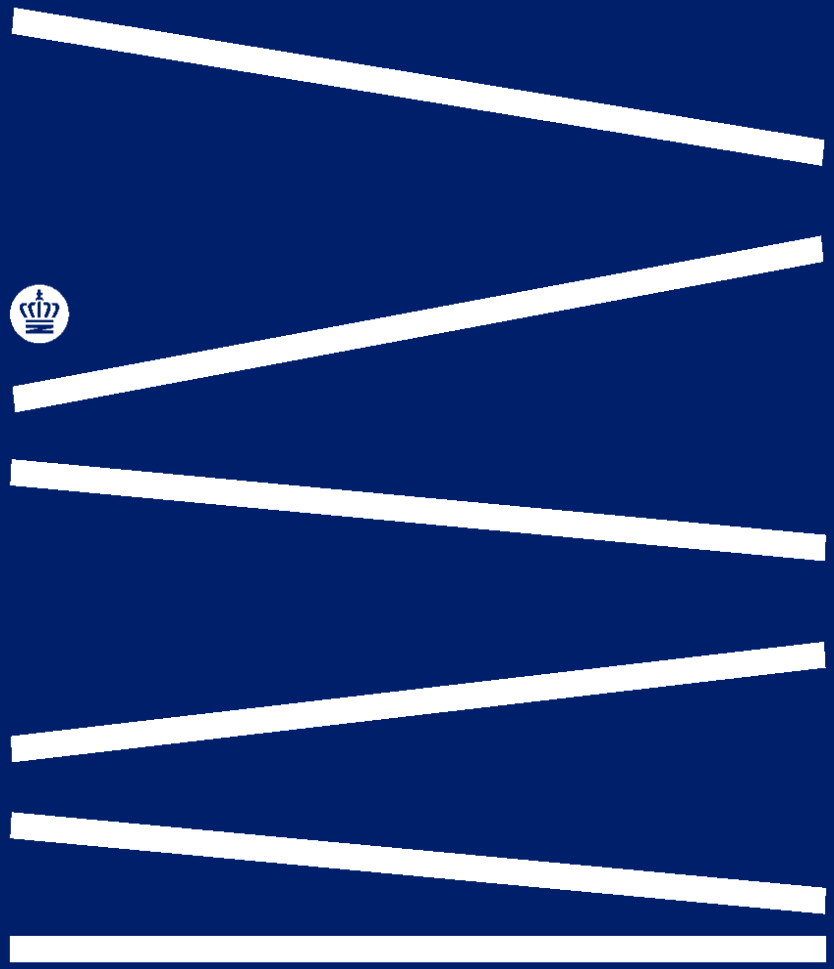
Implications

— The complex nature of the archived web makes the establishing of a corpus within a web archive a very complex undertaking
— It is imperative for researchers to acquaint themselves with the specific digital nature of the archived web in general as well as with the characteristics of the web collection that is studied
— Not all is documented => continuous dialogue with curators and IT developers
—Documentation of all the choices made during the corpus creation in order to reproduce the algoritm and its results

ESTABLISHING A CORPUS OF THE ARCHIVED WEB

The case of the Danish web from 2005 to 2015

Niels Brügger, Aarhus University, DK
Ditte Laursen, Royal Danish Library
Janne Nielsen, Aarhus University, DK
20190606 IIPC, Zagreb, Croatia



DET KGL. BIBLIOTEK
Royal Danish Library

AARHUS UNIVERSITY

The historic context of web archiving and the web archive: reconstructing and saving the Dutch national web using historical methods
Kees Teszelszky (presented by Kees)

Towards a national web archive in a federated country: a Belgian case study
Sally Chambers, Peter Mechant, & Friedel Geeraert (presented by Friedel)

The curious case of archiving .eu
Helen Hockx-Yu, Ditte Laursen, & Daniel Gomes (presented by Daniel)

Exploring the "French Web" of the 1990s
Valérie Schafer (presented by Valérie)

Establishing a corpus of the archived web: the case of the Danish web from 2005 to 2015
Niels Brügger, Ditte Laursen, & Janne Nielsen (presented by Ditte)