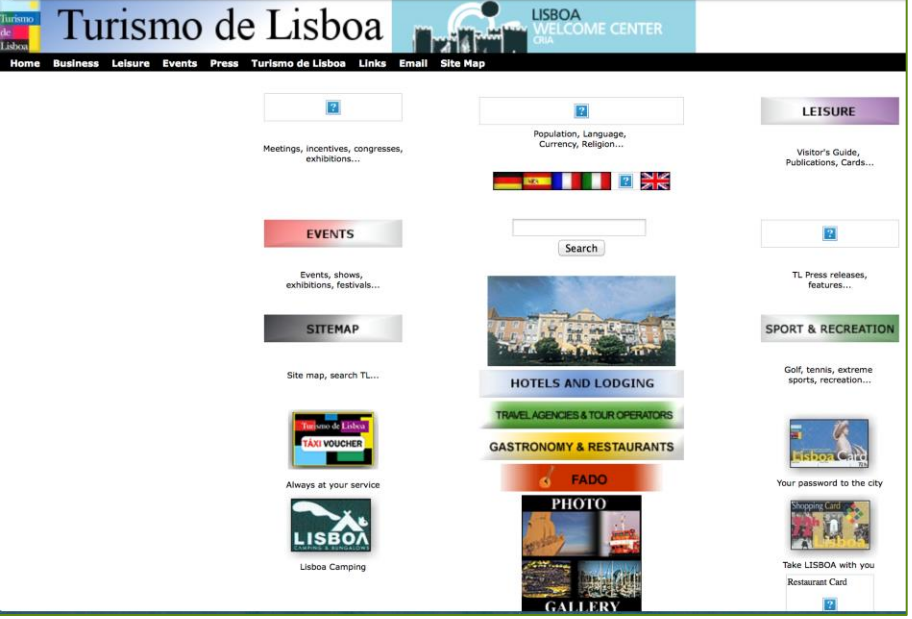
AARHUS
UNIVERSITY

13 NOVEMBER 2014

The web, todays ephemera — web archives and research infrastructure

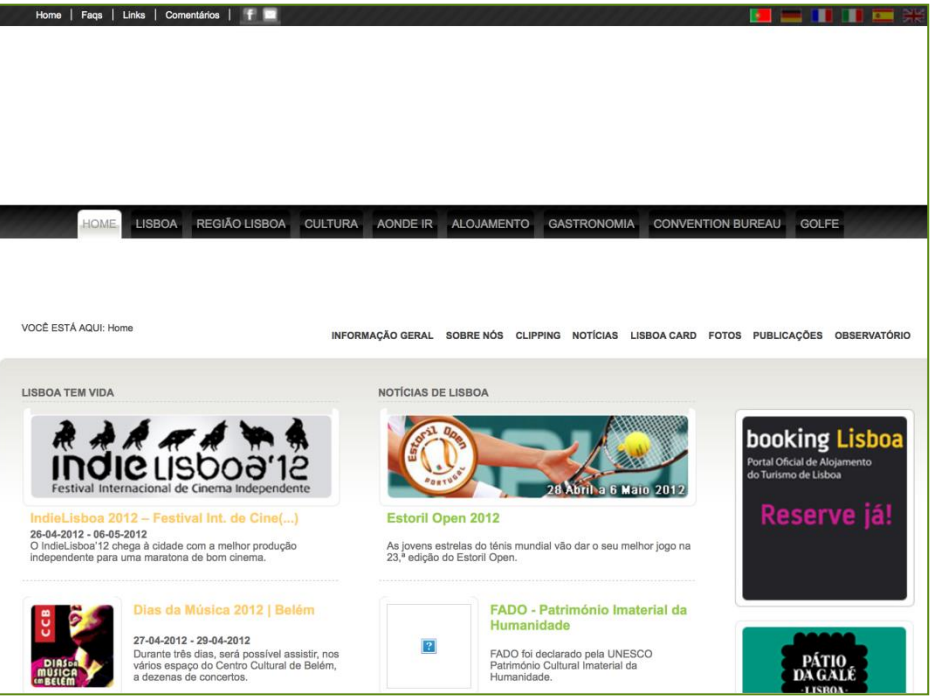
Niels Brügger
HEAD OF NETLAB &
OF THE CENTRE FOR INTERNET STUDIES

UNIVERSITET



visitlisboa.com — 2003, the Internet Archive

visitlisboa.com — 2006, the Internet Archive



visitlisboa.com — 2012, the Internet Archive

AARHUS
UNIVERSITY

Lessons to be learned?

The importance of the web is growing

More and more of our societal, cultural, political, etc. communication take place on the web

The web of the past disappears

40% changed, 40% removed, 20% still there after one year

If we want to document the present or study the past on the web we have to archive it

'We' can be a scholar/group of scholars or a (trans)national web archive such as the Internet Archive or Netarkivet

Web archiving matters for anyone who wants to use the web as a source in any kind of study

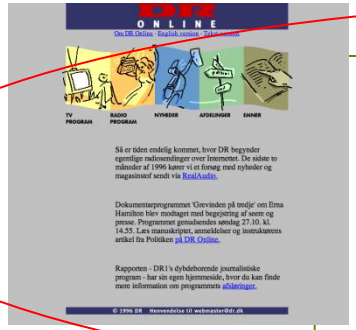
1. Introduction
   - Digitized, Born-digital, and Reborn-digital Materials
   - Digital Humanities — a systematic approach
2. The challenges of the archived web
   - The characteristics of the archived web
   - Analytical and methodological consequences
3. NetLab
4. RESAW — a REsearch infrastructure for the Study of Archived Web materials
5. The obligation to be remembered?

Digitized

Previously analog material which has been digitized.

Born-digital

Has never existed in any other form than digital.



Reborn-digital

Born-digital material which has been collected and preserved, and which to some degree has been changed in this process.

2. THE CHALLENGES OF THE ARCHIVED WEB

The characteristics of the archived web

Web archiving

- Different archiving purpose and strategy
- Differences as to technological choices

Macro web archiving

- archiving institutions, such as national libraries
- aiming at preserving the cultural heritage of, for instance, a nation state
- allows for as many different kinds of research projects as possible in the future

Micro web archiving

- individual scholars or groups
- in relation to, for instance, a specific research project
- usually calibrated to fit the research project in question

The web archive is a real-time archive

- The online web is changed or deleted with an unprecedented pace
- Must be collected and archived here and now, while it is still online

The archived web is a reborn, unique and deficient version and not simply a copy of what was once online

- The archiving institution that wants to archive the online web must make a number of choices
- What is archived is almost never a copy on a 1:1 scale of what was once online
- A collection of unique versions which did not exist before the act of archiving
- It is created in and by the process of archiving, which is why it can be considered 'reborn' digital material

An actively created and subjective re-construction

- subjective because choices have to be made between different archiving forms and strategies (made by either an individual or an institution)
- a re-construction in the sense that it is re-created on the basis of a variety of archived web elements that are re-assembled and re-combined in the archive

Thus, the archived web document is the result of an active process and in this sense it does not exist prior to the act of archiving.

Almost always deficient — for two reasons

- technical reasons (soft- or hardware), for instance words, images, graphics, sounds, moving images can be missing, or some of the possibilities of interaction can be non-functional in the archived web document

- the dynamics of updating, that is the fact that the web content might have changed during the process of archiving, and we do not know if, where, and when this happens — an example

”During the Olympics in Sydney in 2000, I wanted to save the website of the Danish newspaper *JyllandsPosten*. I began at the first level, the front page, on which I could read that the Danish badminton player Camilla Martin would play in the finals a half hour later.

My computer took about an hour to save this first level, after which time I wanted to download the second level, ’Olympics 2000’. But on the front page of this section, I could already read the result of the badminton finals (she lost).

The website was — as a whole — not the same as when I had started; it had changed in the time it took to archive it, and I could now read the result on the front page, where the match was previously only announced.”

N. Brügger: *Archiving Websites*, 2005, pp. 22-23

Consequences

- we cannot be sure that we have everything in our archive — we will always have lost something in the asynchronous relationship between updating and archiving

- we are also in danger of getting something that in a way was never there — something that is different from what was really there

The process of archiving

- creates a unique version and not a copy
- a version of an original which we can never expect to find in the form it actually took on the web
- neither can we find an original among the different versions, nor can we reconstruct an original based on the different versions

Something is missing

- The web material is incomplete compared to what was once online — two general types of incompleteness
  - The user of a web archive will miss some of the information about the web which is usually at hand on the online web
  - Individual web elements and possibilities of interaction may be missing

Something is missing

What is specific for the incompleteness of web archives is not that things are missing, but rather that they may be missing in ways which make it very difficult to determine if something is missing at all as well as what and where

- No stable original to compare with
- Incompleteness is rarely documented

2. THE CHALLENGES OF THE ARCHIVED WEB

Analytical and methodological consequences

Hyperlinks become inconsistent

- The structure of hyperlinks is an integrated part of the archived web and not just an added feature of the archive

- Gives rise to problems of inconsistency related to time and space
  - Temporal inconsistency between the link source and the link target
  - Spatial inconsistency if the link target is not archived at all

- Difficult to determine if — and to what extent — the archived web material is inconsistent or not

The archived web is edited and editable

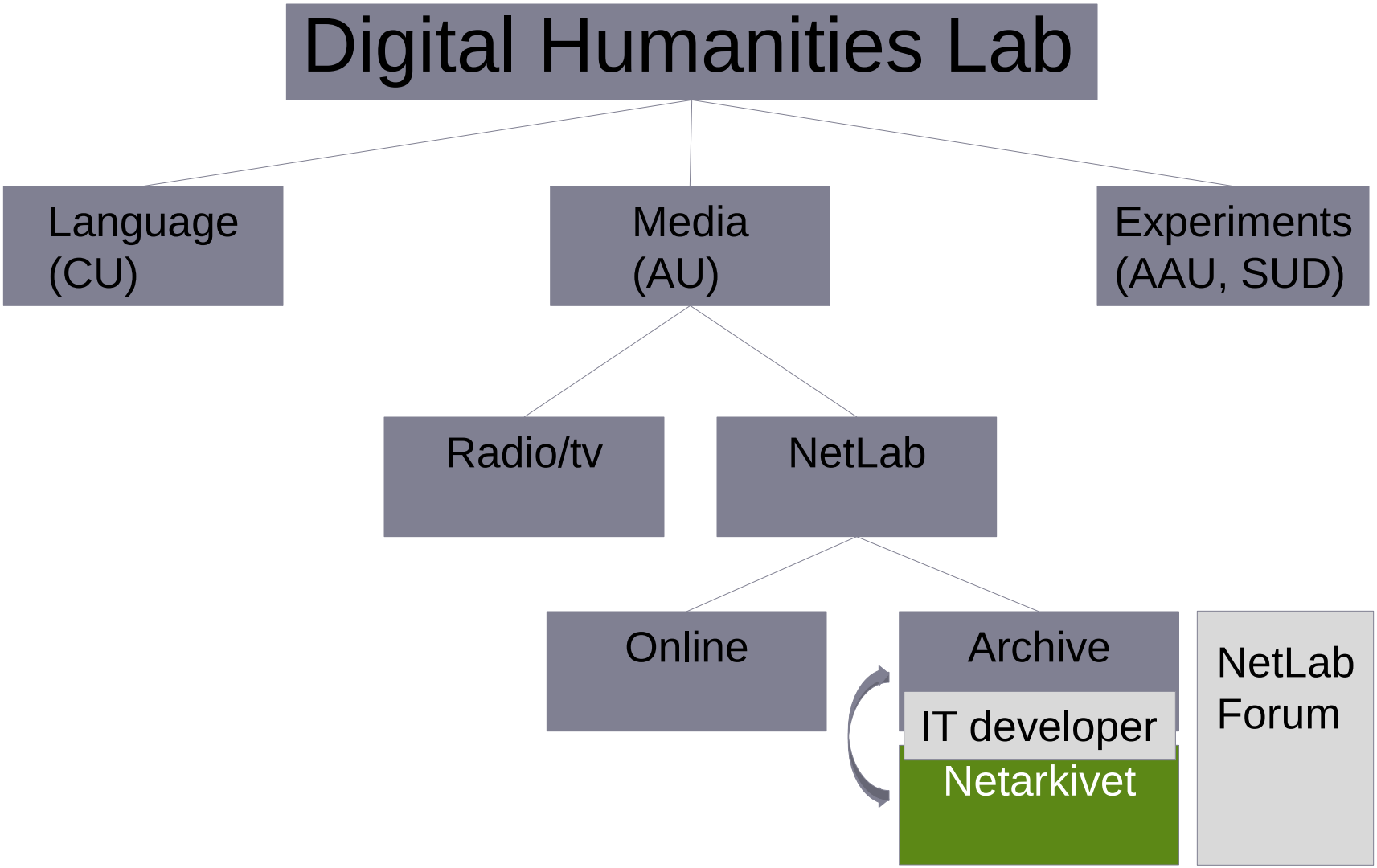
- The archived content itself as well as the division of the material in elements can be changed
- Any 'montage' of the archived elements in the archive — or any extraction from the archive — is also an editing of these elements
- Reason: the subdivision of the archived material and the subsequent combination of elements are not necessarily inscribed in the material itself
- A continuum with no clear-cut temporal or spatial subdivisions inscribed by the producer; the subdivisions are editable, scalable, and random, and they are made a posteriori by either the web archive or the scholar

Subsequent processing is lacking

- The amount and the complexity of the archived material do not allow for systematic and detailed processing of the entire archive once the web has been archived (exceptions exist)
- Have to make do with either the metadata provided by the archived web itself (e.g. meta-tags in the source code), or with the log files from the archiving process, if the archive makes them available

3. 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
Archive
IT developer
Netarkivet
NetLab Forum

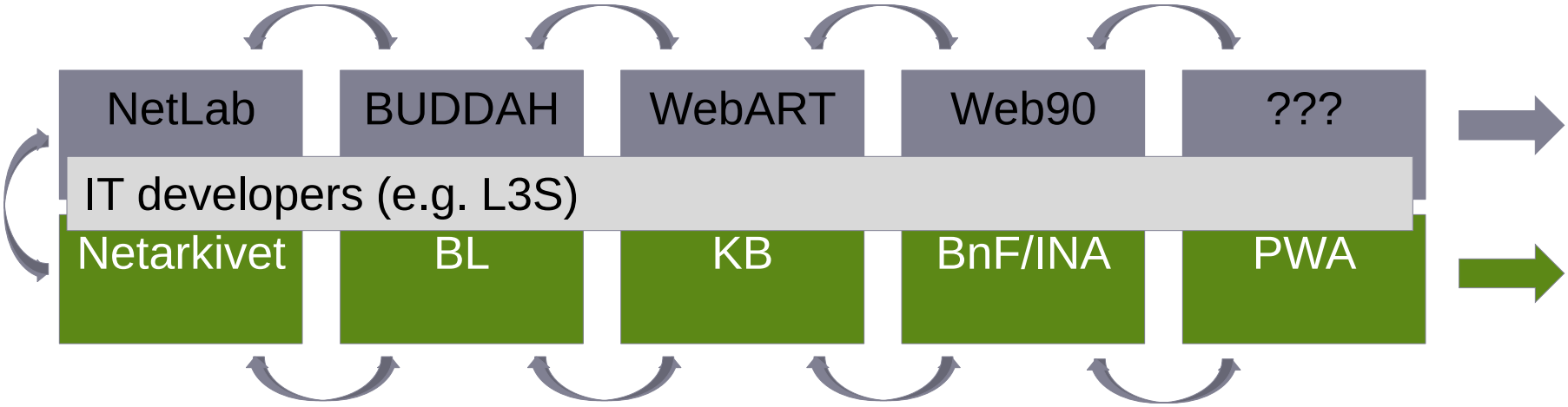
4. RESAW

'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

4. RESAW


NetLab
BUDDAH
WebART
Web90
???
IT developers (e.g. L3S)
Netarkivet
BL
KB
BnF/INA
PWA

4. RESAW

- A larger network of relevant institutions and researchers, European as well as international (app. 40 participants)
- 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'

4. RESAW

The group coordinates a number of activities in 2014-15, including:

›a seminar, London, 3-4 December 2014

›an international conference, Aarhus, 8-10 June 2015 entitled 'Web Archives as scholarly Sources: Issues, Practices, and Perspectives' — call out soon

›small pilot projects (e.g. how the internet domain .eu can be archived, Eurovision Song Contest...)

›resaw.eu

5. THE OBLIGATION TO BE REMEMBERED?

› The right to be forgotten revisited — another agenda with web archives

5. THE OBLIGATION TO BE REMEMBERED?

- Web archives are in general not *archives* strictly speaking
- Emerges out out librarian world

5. THE OBLIGATION TO BE REMEMBERED?

The major difference regarding what cultural heritage institutions collect and preserve:

- Objects are collected and preserved by museums
- Documents are collected and preserved by:
  - Libraries if published
  - Archives if not published

5. THE OBLIGATION TO BE REMEMBERED?

’Publication’ ≠ ’printed’

The core of the concept of ’publication’:
›the word ’public’
›in line with Immanuel Kant’s ideas of *Publicität*
’publication’ means ’made available to the public’
›regardless of which media type is used for doing this

5. THE OBLIGATION TO BE REMEMBERED?

What is published is that to which every member of the public can get access.

The Danish national web archive Netarkivet is allowed to collect web material behind passwords if every member of the public can get a password, either by simply requesting one or by paying for it

5. THE OBLIGATION TO BE REMEMBERED?

Two consequences:
›Terminology: since a web archive archives the published web, and it emerges out of the library it should rightly be called ‘webrary’
›Collecting and preserving: Whatever is published on the web must be remembered, just as we do with other public media types

5. THE OBLIGATION TO BE REMEMBERED?

The challenges with the web — blurred boundaries — material that is publicly available on the web, but should not have been there:

›published by mistake — social security number, health information, unintentionally on the websites of municipalities or hospitals

›the author was not aware of the reach of his or her expression — the syndication of expressions on personal social media profiles such as Facebook

5. THE OBLIGATION TO BE REMEMBERED?

- published in the sense it has been made publicly available — should thus be preserved within the realm of libraries
- belongs to the private sphere since was not intended for publication, and may then belong to the realm of archives, if should be preserved at all

But I consider the latter an exception, and it must be dealt with in the research process — a matter of research ethics

5. THE OBLIGATION TO BE REMEMBERED?

In conclusion

›Starting point: utterances made available to the public on the web are subject to an obligation to be remembered — and not to a right to be forgotten

›Blurred boundaries must be dealt with in the research process, not in the archiving process — a matter of research ethics

If not we will be unable to write our own history in just a few years time