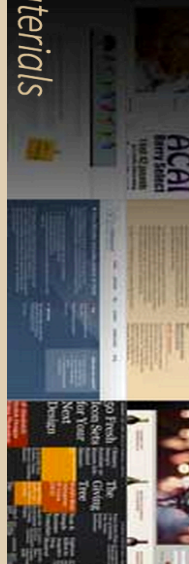
.eu pilot project

Daniel Gomes, Portuguese Web Archive
Helen Hockx-Yu, UK Web Archive
Ditte Laursen, Danish Netarchive

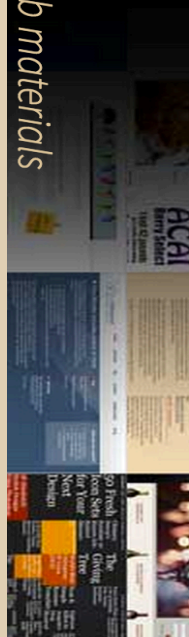
RESAW

A Research infrastructure for the Study of Archived Web materials



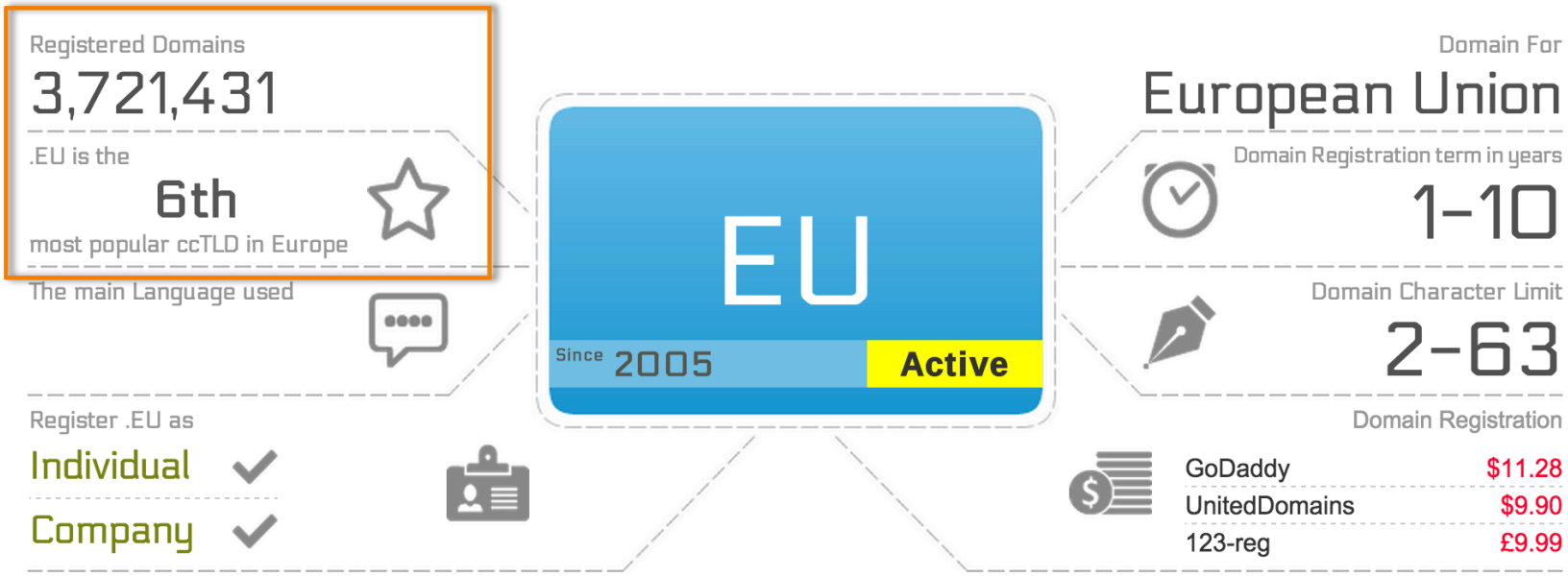
Background

- National responsibility for national domains (ie. .uk, .fr, .dk …)
- .eu covers multiple nations and does not map to any archive's responsibility or mandate
- Some archives have captured some .eu content
  - the Danish net archive: at least 1063 .eu domains
  - UK Web Archive: at least 2812 .eu domains
  - Portuguese Web Archive: at least 2264 .eu domains



Background

- Approx. 3.7 millions .eu registrations (EURid 2012 + domaintyper 2014)
- Without an .eu collection we will have a significant gap in the history of the web and this will impact future generations' understanding of European nations on the web



RESAW

A Research infrastructure for the Study of Archived Web materials

Funding models

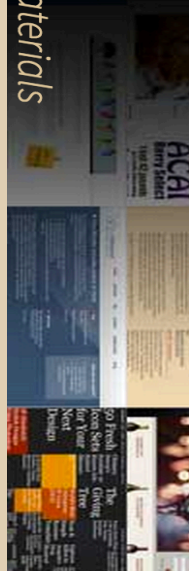
- European Commission on an on going basis
- RESAW consortium, membership fees
- Partial funding from IIPC
- Goodwill of the different archives
- Commercial venture

Collecting .eu

- Full coverage based on .eu domain registrations from EURid
- Selective coverage
  - collecting .eu content from existing web archives
  - individual archives undertake different archiving responsibilities in relation to .eu, coordinated by RESAW
  - RESAW offers a professional service to content owners using .eu domains

RESAW

A Research infrastructure for the Study of Archived Web materials

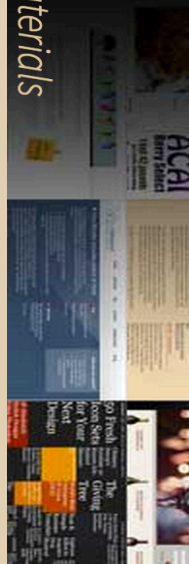


Other issues

- Crawling services
  - Outsourcing to a service provider (ie. Archive-It (Internet Archive), Archivethe.Net (Internet Memory Foundation), DWA/The Web Archive (OIA) …)
  - Each archive uses its own services
  - Development of personal archiving tools
- Governance
  - Collection setup and management
    - start/end dates for one harvest, or ongoing collection effort?
  - Staffing
    - content development group?
    - team of contributing curators?
    - collection administrator?
    - access development group? (legal issues, collection-level metadata, multiple-language support etc.)
- Access

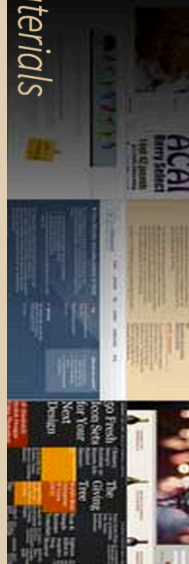
RESAW

A Research infrastructure for the Study of Archived Web materials



Pilot Crawl 2014-15

- Purpose: To obtain a better understanding of the domain
- 34 138 unique sites (identified from Google, DMOZ, domaintyper.com, Alexa.com, etc.) + new ones identified during the crawl
- Limits: 3 levels, 1000 URLs per site, 100 million URLs
- Storage limit: 4 TB
- Indexing using PWA software
- Search prototype with restricted access
- Crawl available open access at archive.pt, Feb 15th



Possible benefits

- Research: Providing new opportunities
- Access: content that archiving institutions cannot make publicly available from their own servers because of legal issues or the need to obtain permissions may be made available through a vendor host
- Collaboration: formalizing and supporting collaboration, across a wider number of institutions, and preventing over-dependence on individual institutions
- Profiling: Enabling higher-profile promotion and advocacy for RESAW and for web archiving in general

Possible risks

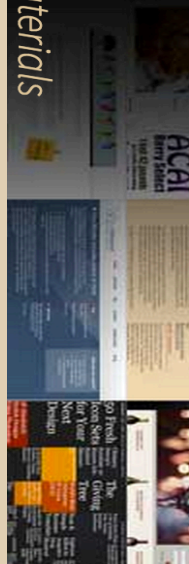
- That RESAW ownership of the collaborative collection and responsibility for long-term preservation is not sustainable
- Legal risks: bypass legal deposit laws, risks with determining legal responsibility, and risks with the variety of legal requirements across many countries

Relations to consider

- IIPC's collaborative collection building initiative
- RESAW collaborative collection building initiative (ie. Eurovision Song Contest)
- National archives (parallel collection building)
- European Digital Library

RESAW

A Research infrastructure for the Study of Archived Web materials

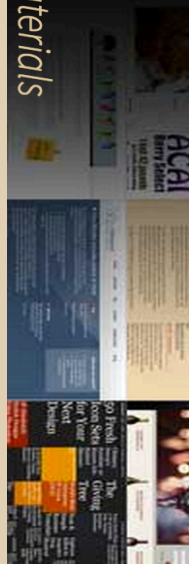


.eu pilot project

Daniel Gomes, Portuguese Web Archive
Helen Hocks-Yu, UK Web Archive
Ditte Laursen, Danish Netarchive

RESAW

A Research infrastructure for the Study of Archived Web materials



Questions to consider

- Full coverage vs. selective coverage
- Separate collection or collection based on archives existing collection practices
- Outsourcing or