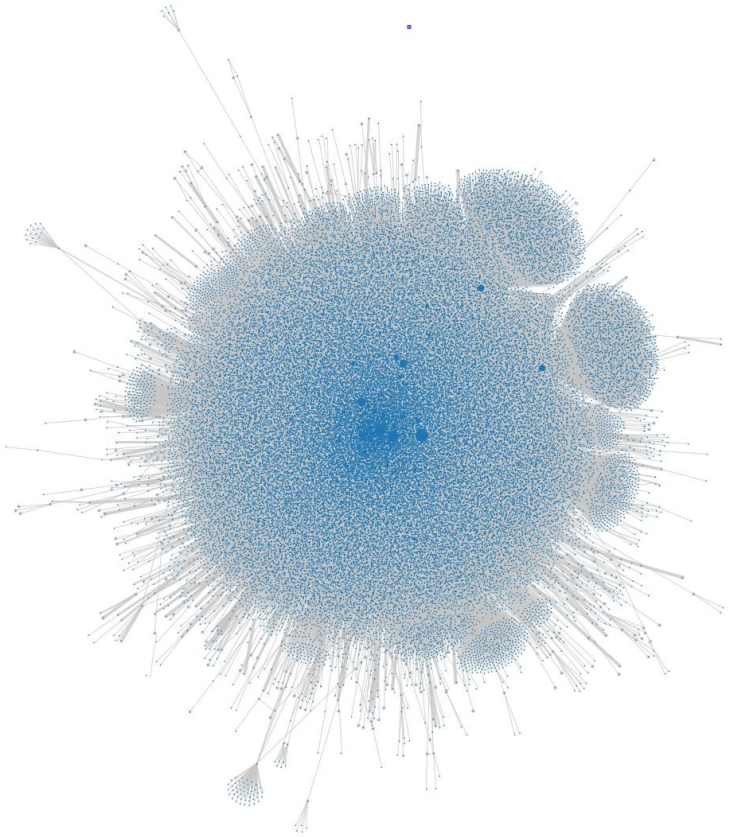
Using Open Data To Explore The UK Web

Andrew Jackson

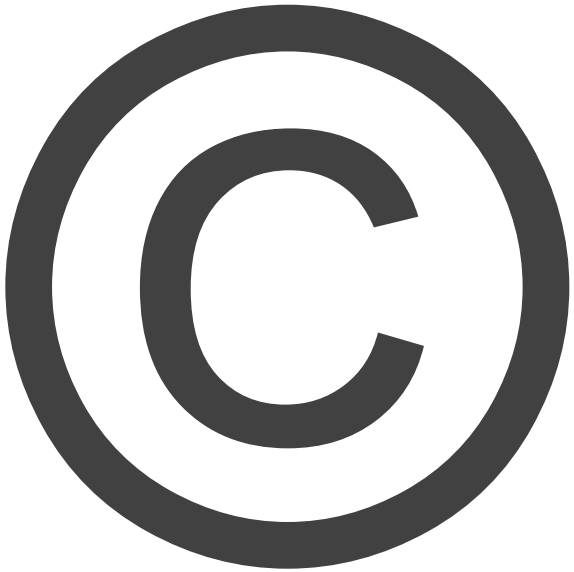
UK Web Archive Technical Lead

The UK Web Archive

- Three collections:
  - Selective Archive (since 2004)
  - Legal Deposit Archive (since 2013)
  - JISC Historical Archive (1996-2013)
- Statistics:
  - Over six billion resources
  - Over 100TB compressed data
- Goals:
  - Preserve UK web history
  - Support access
  - Enable research

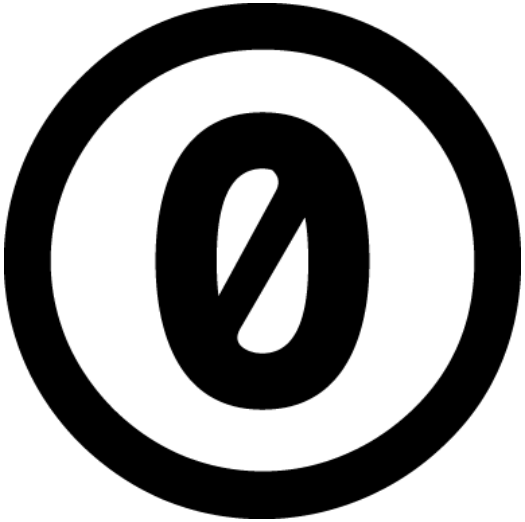


The Problem



The Tactic

- Secondary Datasets:
  - Composed of facts about the content
  - But not ‘substitutable’ for the content
- Part of a long-standing tradition:
  - The British Library’s bibliographic data has always been openly accessible
- Probably not copyrightable:
  - Released as CC0 to avoid any ambiguity



The Datasets

- JISC UK Web Domain Dataset (1996-2013):
  - Format Profile
  - Geo-index
  - Host-level Links
  - Crawled URL Index
- UK Selective Web Archive:
  - Website Classification Dataset
- Hosted at:
  - http://data.webarchive.org.uk/opendata/

Host-to-Host Links (1996-2010)

- Simple text file documenting links:

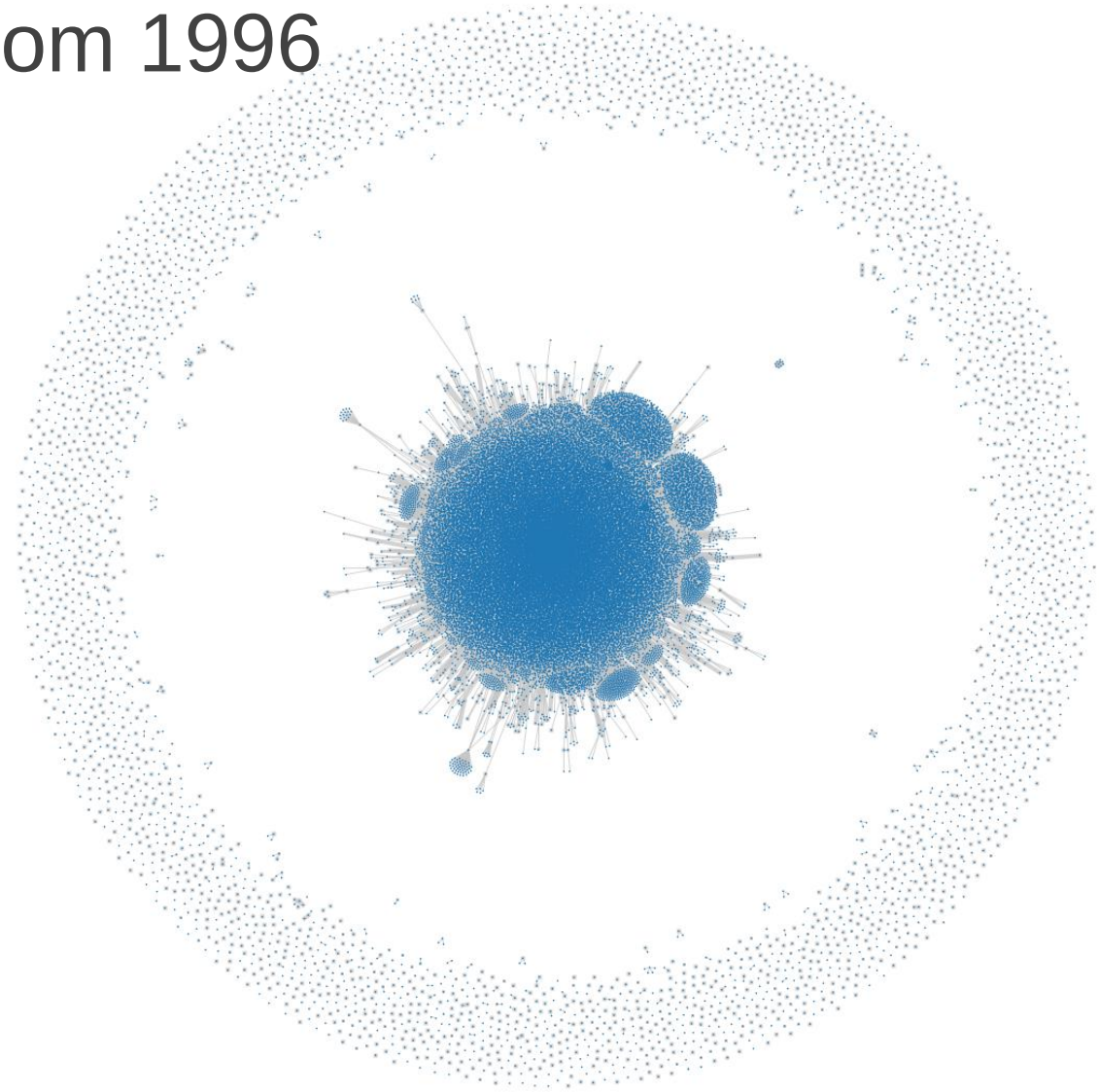
1996|appserver.ed.ac.uk|portico.bl.uk 1
1996|art-www.acorn.co.uk|portico.bl.uk 1
1996|astra.ich.ucl.ac.uk|portico.bl.uk 1
1996|back.niss.ac.uk|portico.bl.uk 1
1996|beta.bids.ac.uk|portico.bl.uk 2
1996|blaiseweb.bl.uk|blaiseweb.bl.uk 4
1996|bonsai.iielr.dmu.ac.uk|portico.bl.uk 4

- Large (19GB when compressed) but easy to filter:

grep ".bl.uk" host-linkage.tsv > bl-uk.tsv

grep ”.ac.uk" host-linkage.tsv > ac-uk.tsv

Links From 1996



The Future

- What facts can we extract for you?
  - Very happy to work with researchers to see what data they need.
- Get in touch:
  - Email: Andrew.Jackson@bl.uk
  - Twitter: @UKWebArchive
- Get the data:
  - http://data.webarchive.org.uk/opendata/