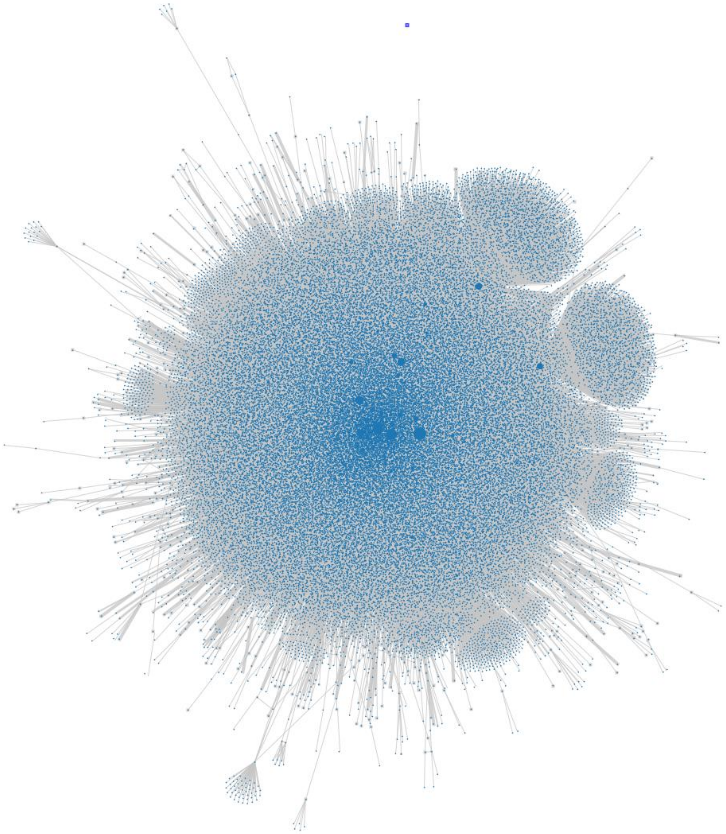
Ten years of the UK Web Archive: What have we saved?

Andy Jackson (@anjacks0n)

UK Web Archive Technical Lead

The UK Web Archive

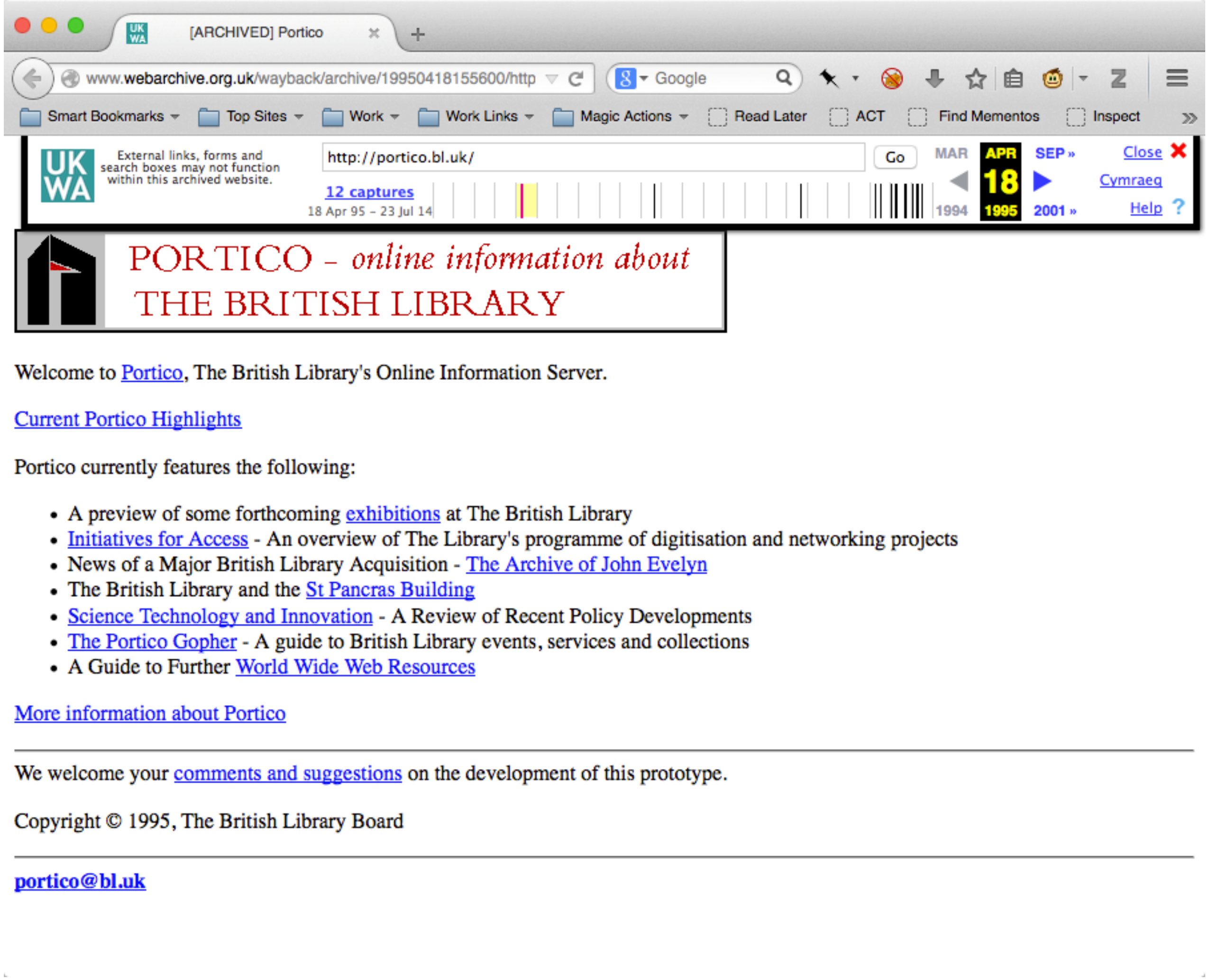
- Three collections:
  - Open Archive (since 2004)
  - Legal Deposit Archive (since 2013)
  - JISC Historical Archive (1996-2013)
- Statistics:
  - Over eight billion resources
  - Over 160TB compressed data
- Goals:
  - Preserve UK web history
  - Support access
  - Enable research



Understanding Our Collections

© Neil Howard
CC-BY-NC
https://flic.kr/p/cUd91m

Resource-Level Access



PORTICO – *online information about* THE BRITISH LIBRARY

Welcome to Portico, The British Library's Online Information Server.

Current Portico Highlights

Portico currently features the following:

- A preview of some forthcoming exhibitions at The British Library
- Initiatives for Access - An overview of The Library's programme of digitisation and networking projects
- News of a Major British Library Acquisition - The Archive of John Evelyn
- The British Library and the St Pancras Building
- Science Technology and Innovation - A Review of Recent Policy Developments
- The Portico Gopher - A guide to British Library events, services and collections
- A Guide to Further World Wide Web Resources

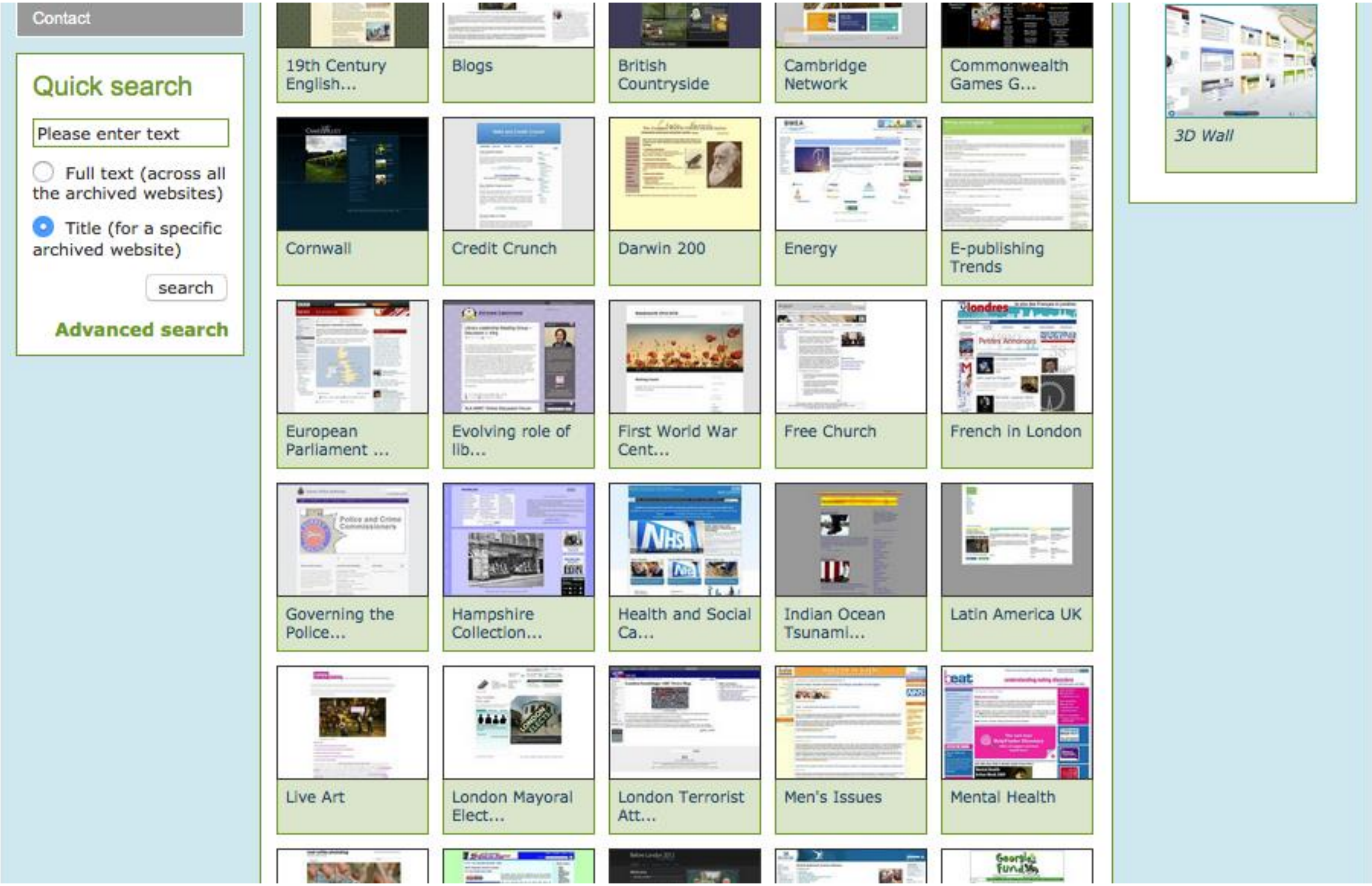
More information about Portico

We welcome your comments and suggestions on the development of this prototype.

Copyright © 1995, The British Library Board

portico@bl.uk

Curated Collections



Quick search

Please enter text

○ Full text (across all the archived websites)

● Title (for a specific archived website)

19th Century English... | Blogs | British Countryside | Cambridge Network | Commonwealth Games G...

Cornwall | Credit Crunch | Darwin 200 | Energy | E-publishing Trends

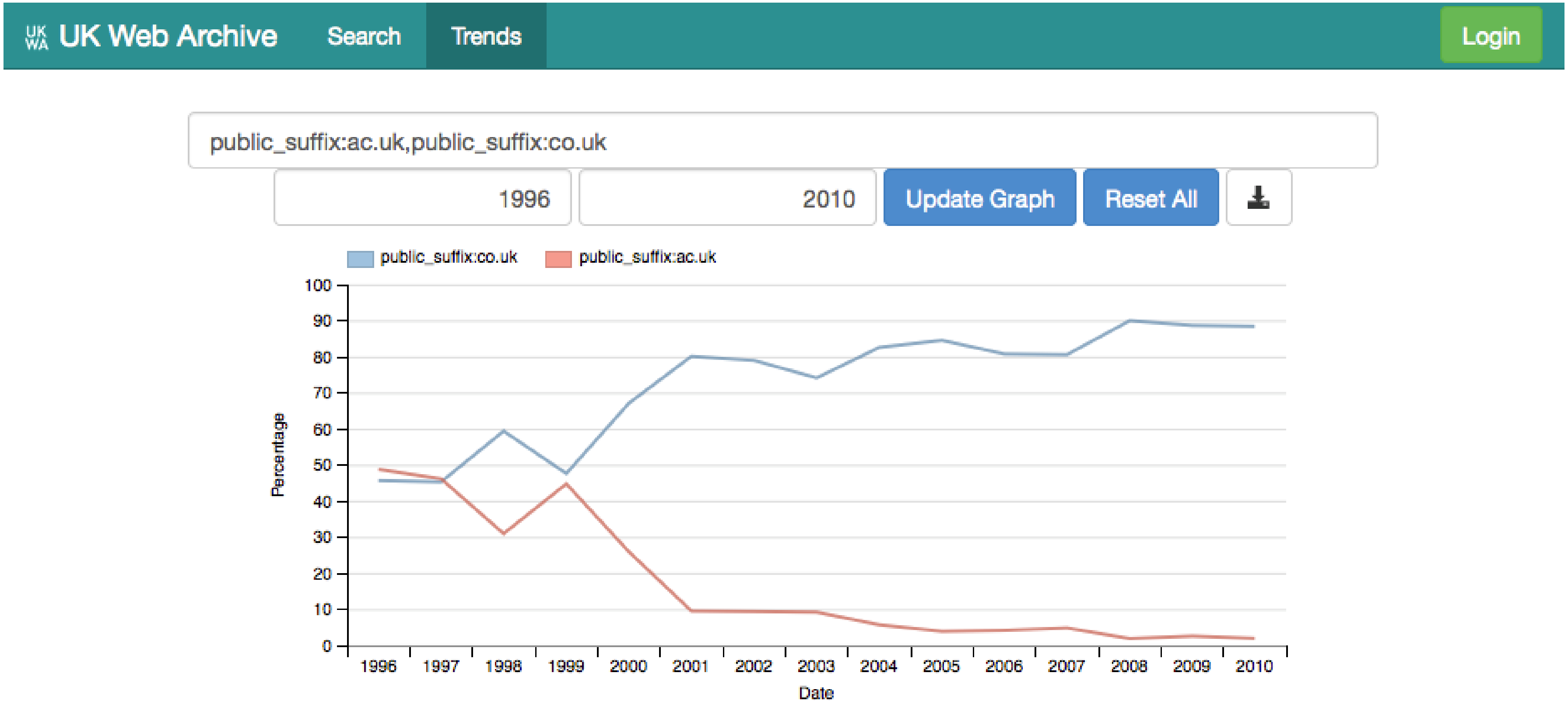
European Parliament ... | Evolving role of lib... | First World War Cent... | Free Church | French in London

Governing the Police... | Hampshire Collection... | Health and Social Ca... | Indian Ocean Tsunami... | Latin America UK

Live Art | London Mayoral Elect... | London Terrorist Att... | Men's Issues | Mental Health

3D Wall

Full-Text Discovery & Trend Analysis



public_suffix:ac.uk,public_suffix:co.uk

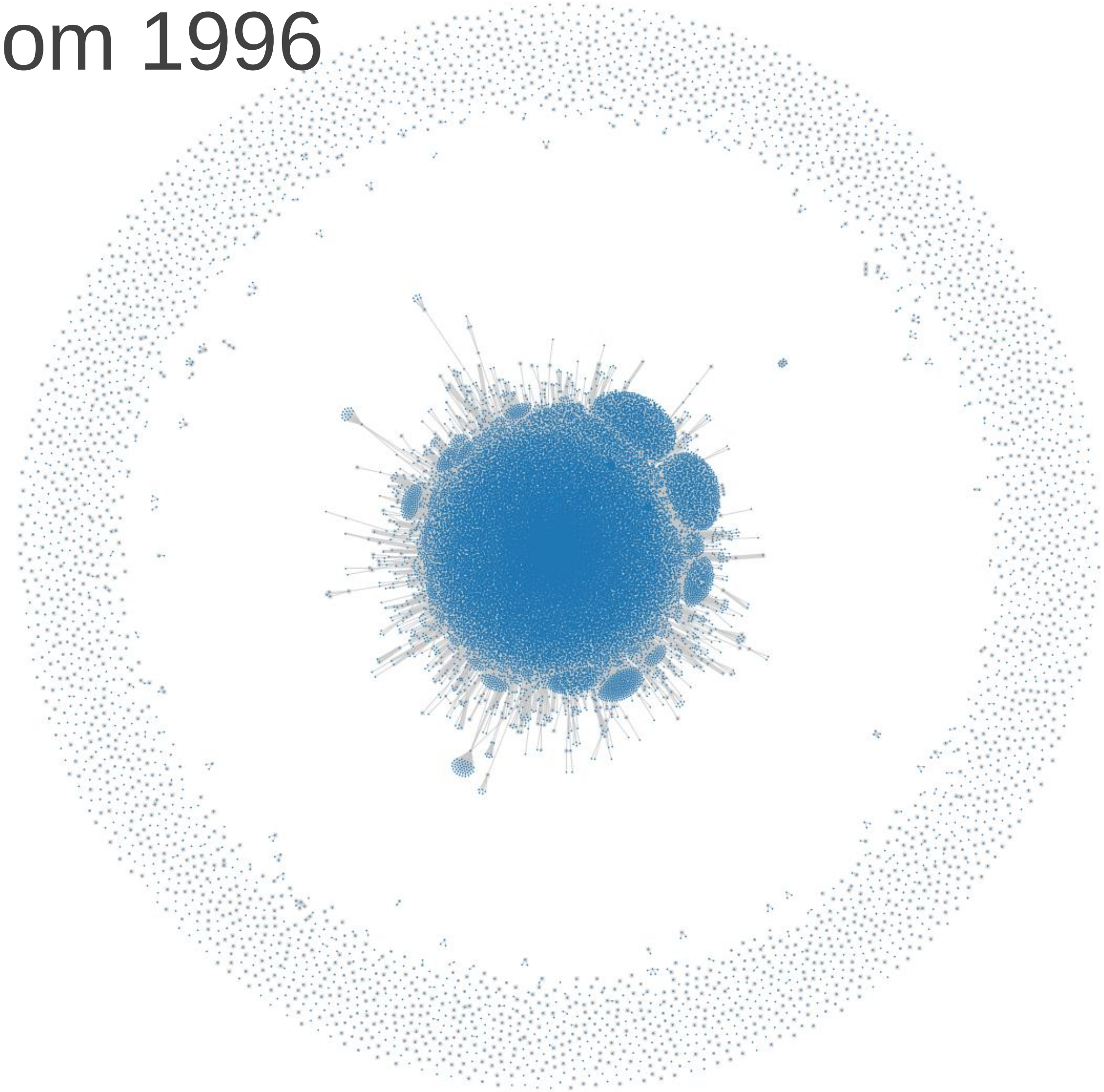
1996

2010

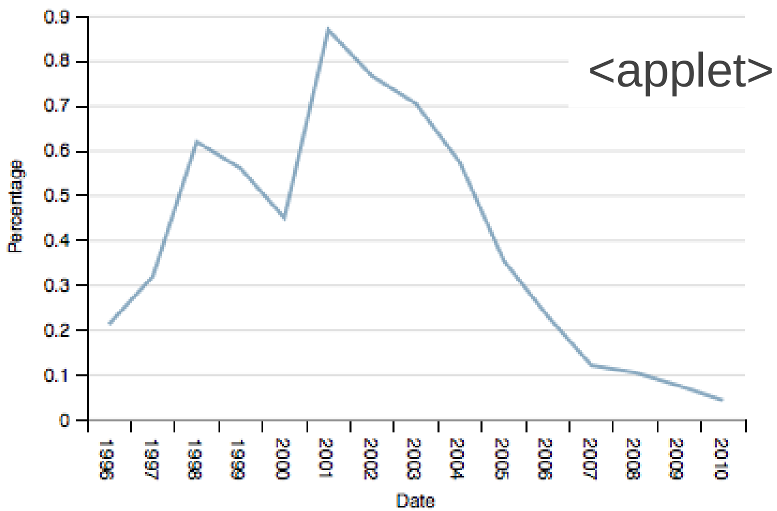
Secondary Datasets

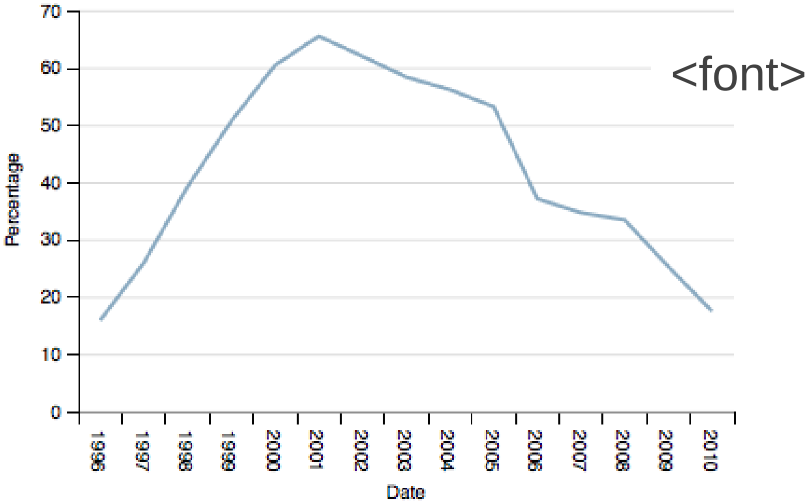
- JISC UK Web Domain Dataset (1996-2013):
  - Format Profile
  - Geo-Index
  - Host-Level Links
  - Crawled URL Index
  - WATs (not released yet)
- UK Open (Selective) Web Archive:
  - Website Classification Dataset
- Available as CC0 downloads:
  - http://data.webarchive.org.uk/opendata/

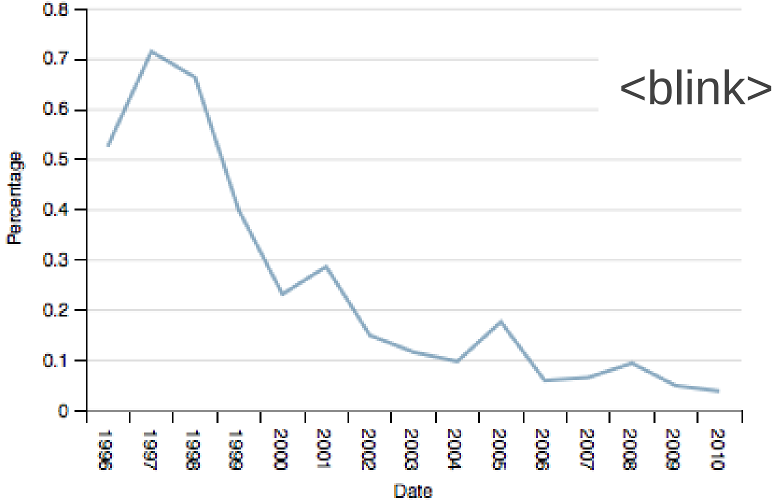
Links From 1996

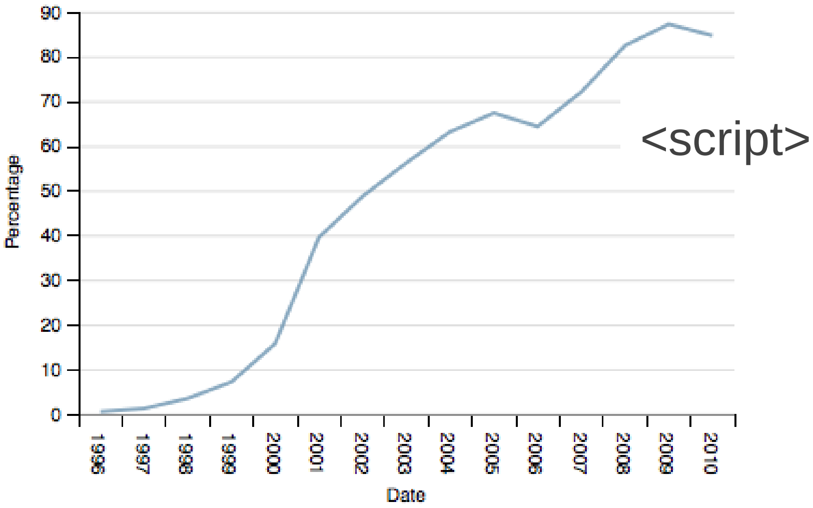


Format & Feature Analysis


<applet>


<font>

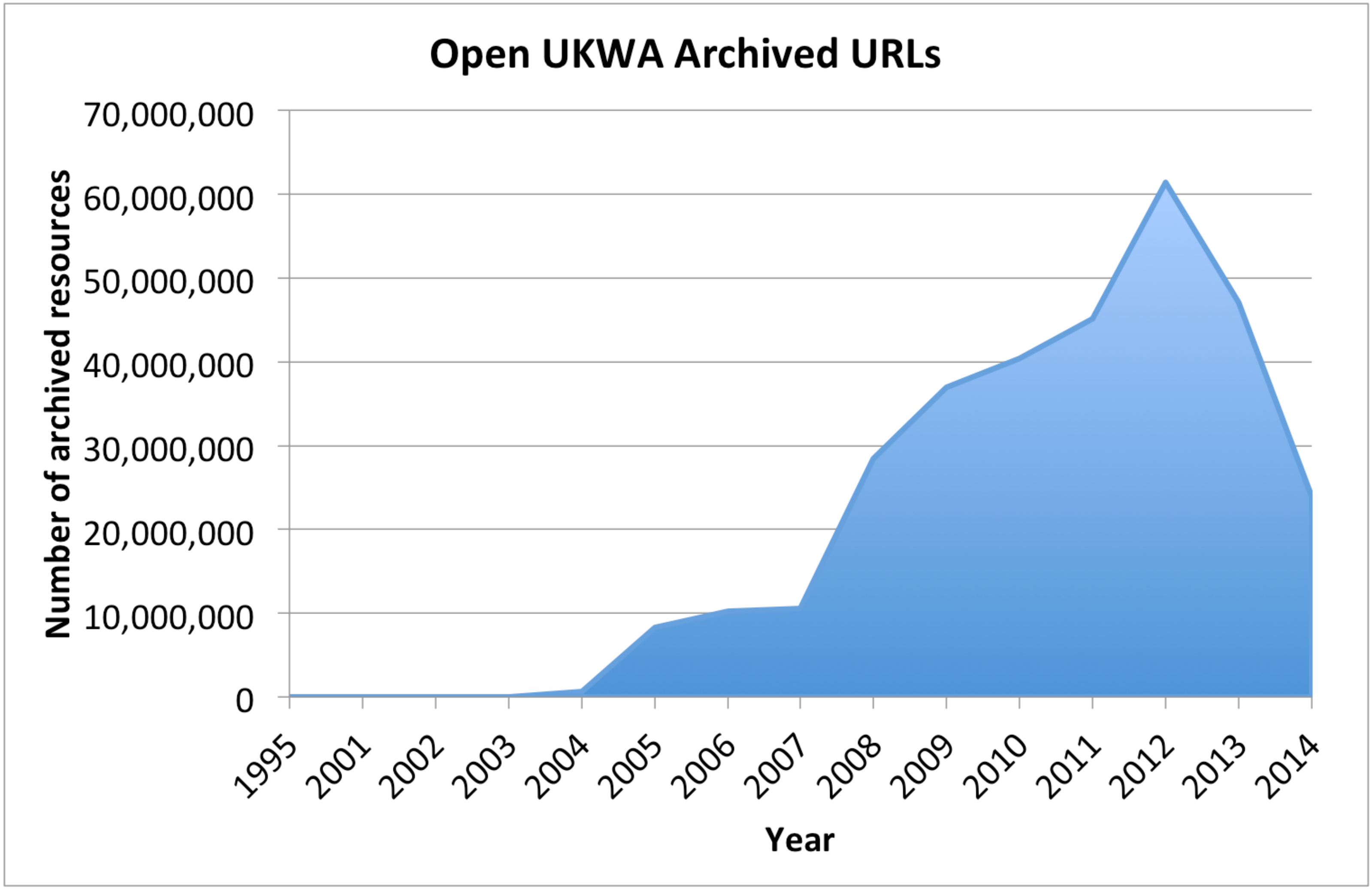

<blink>


<script>

Putting Our Archives In Context

- Looking inward is not enough:
  - To understand the value of our collection, we need to look beyond our walls and put it in context.
- Was it worth archiving?
  - How much of our collection is still on the live web?
  - How bad is reference rot in the UK domain?

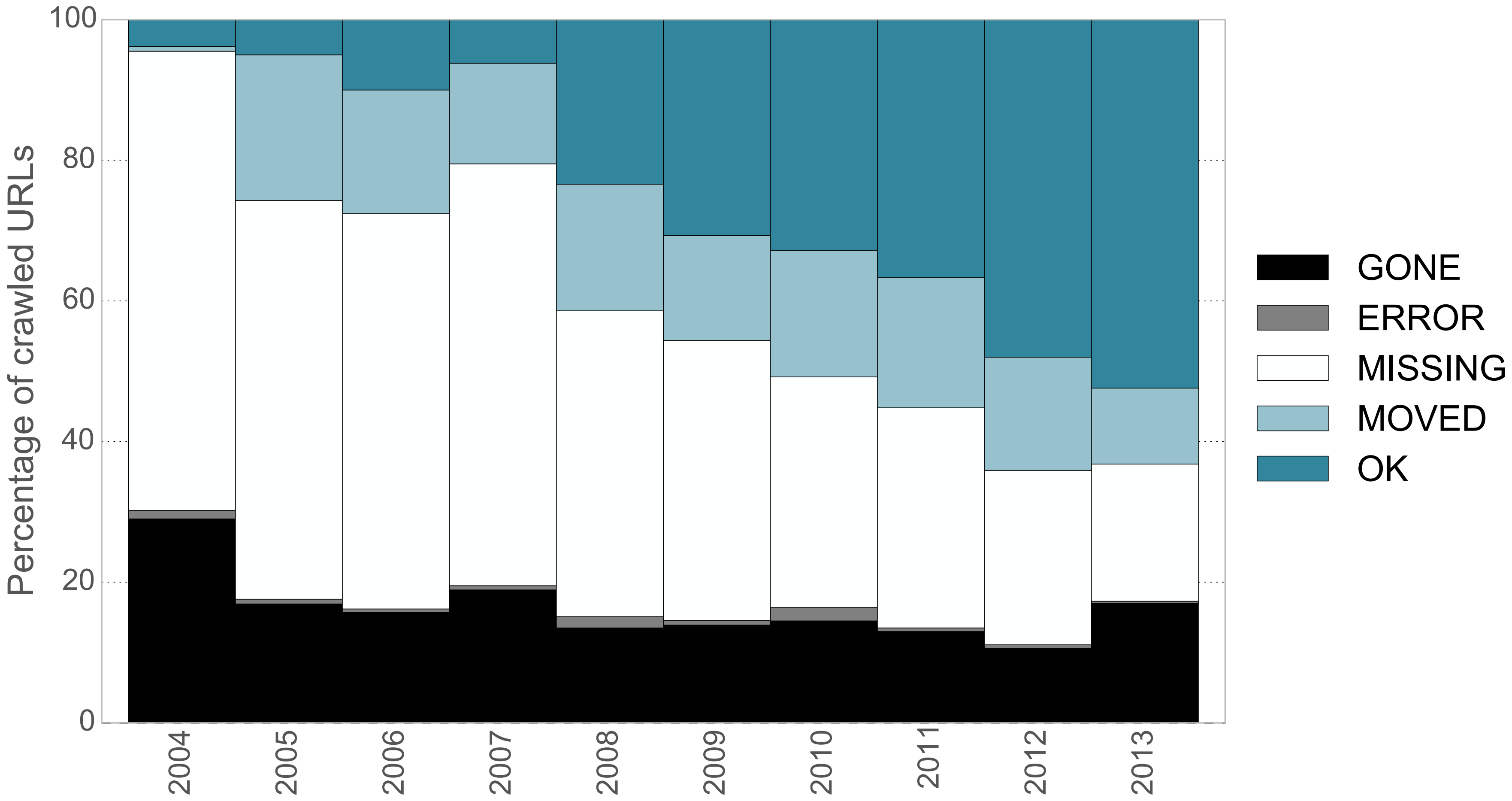
Open UKWA Crawl History



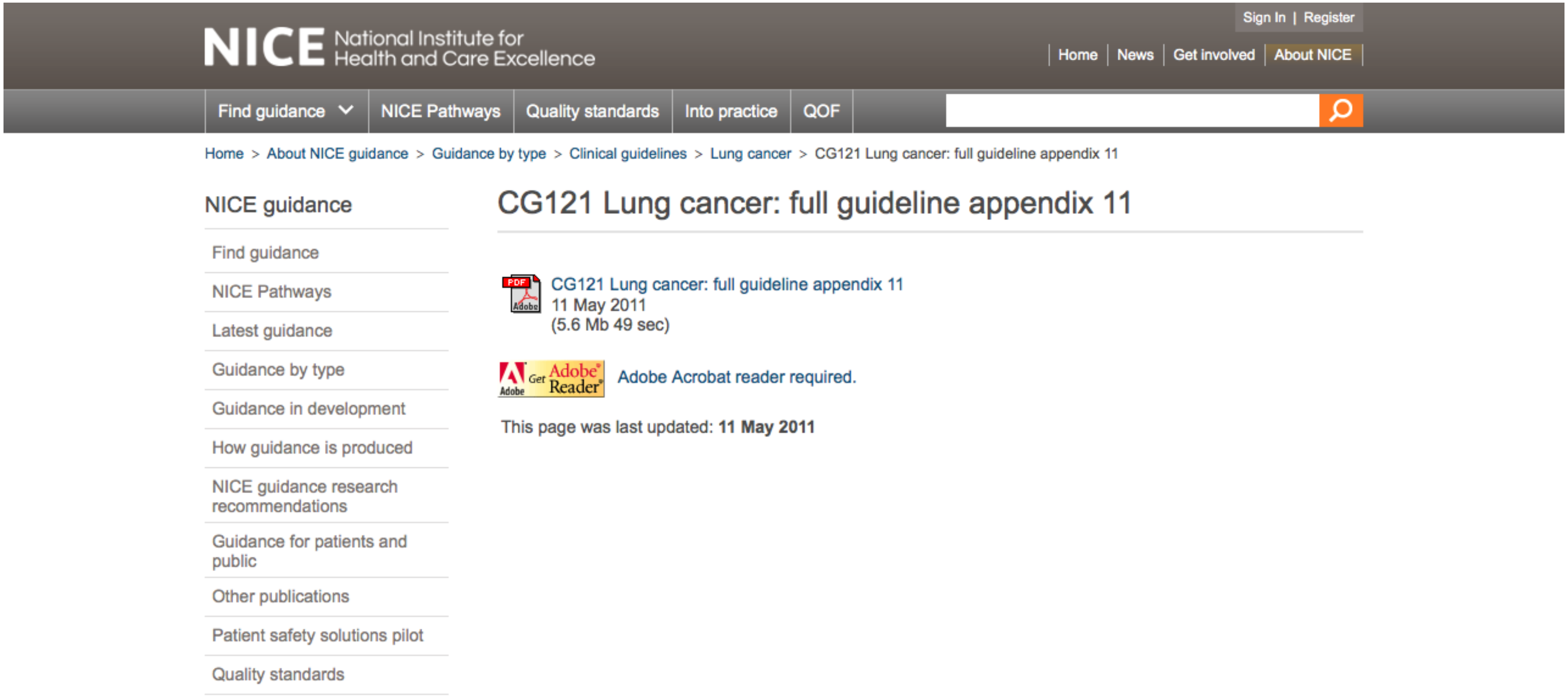
Sampling The URLs

- Use a random sample 1,000 URLs per year:
  - If the host name does not resolve, or is unreachable:
    - **GONE**
  - If the server responds with an error:
    - **ERROR**
  - If the server response leads to *404 Not Found*:
    - **MISSING**
  - If the server response leads to a valid resource:
    - **MOVED** (if via redirects)
    - **OK** (otherwise)
- n.b. ‘soft 404s’ are surprisingly rare (< 1%)

Where Are They Now?



NICE Example



Home > About NICE guidance > Guidance by type > Clinical guidelines > Lung cancer > CG121 Lung cancer: full guideline appendix 11

NICE guidance

- Find guidance
- NICE Pathways
- Latest guidance
- Guidance by type
- Guidance in development
- How guidance is produced
- NICE guidance research recommendations
- Guidance for patients and public
- Other publications
- Patient safety solutions pilot
- Quality standards

CG121 Lung cancer: full guideline appendix 11

CG121 Lung cancer: full guideline appendix 11
11 May 2011
(5.6 Mb 49 sec)

Adobe Acrobat reader required.

This page was last updated: **11 May 2011**

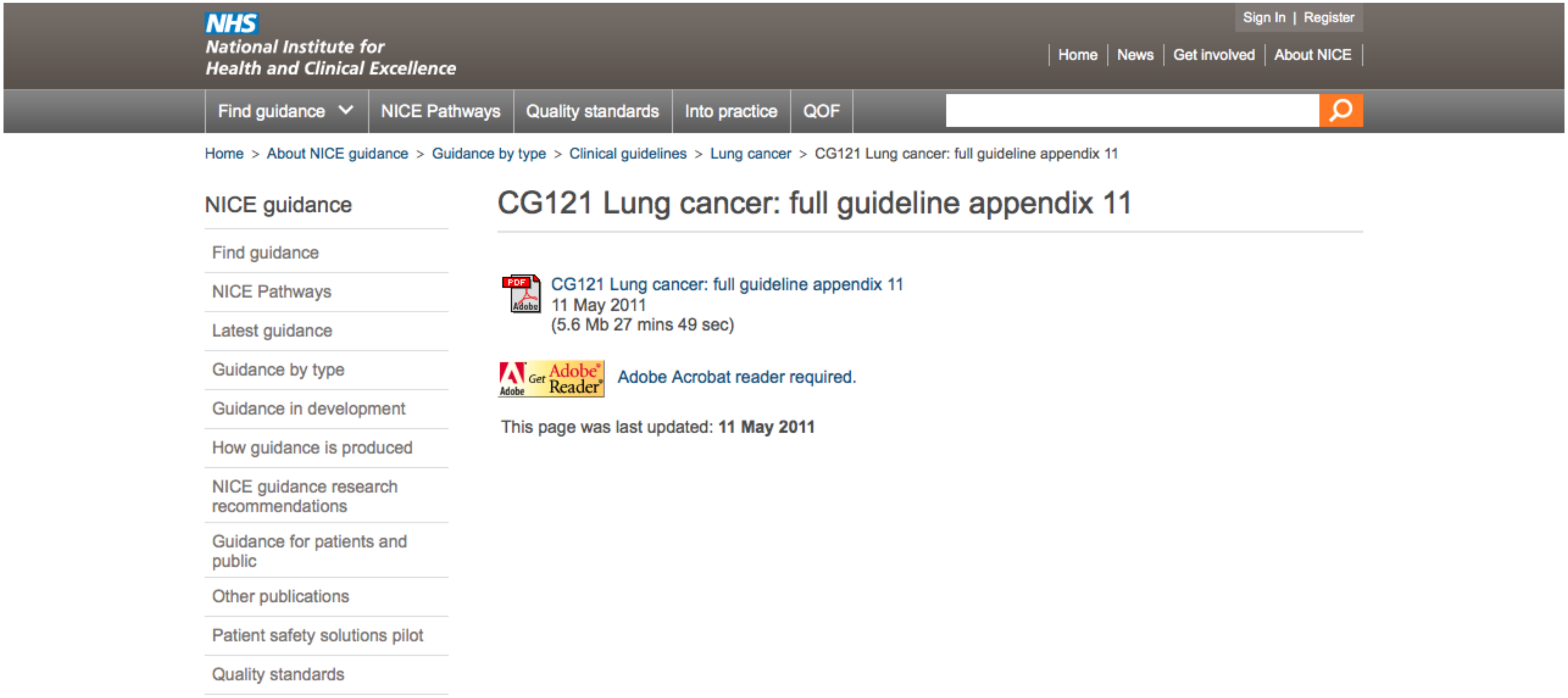
Extract The Text

CG121 Lung cancer: full guideline appendix 11 Sign In | Register Home News Get involved About NICE Find guidance NICE Pathways Quality standards Into practice QOF Conditions and diseases Blood and immune system ? Cancer ? Cardiovascular ? Central nervous system ? Digestive system ? Ear and nose ? Endocrine, nutritional and metabolic ? Eye ? Gynaecology, pregnancy and birth ? Infectious diseases ? Injuries, accidents and wounds ? Mental health and behavioural conditions ? Mouth and dental ? Musculoskeletal ? Respiratory ? Skin ? Urogenital Public health Accidents and injuries ? Alcohol ? Behaviour change ? Cancer ? Cardiovascular disease ? Child health ? Child social care ? Chronic illness ? Diabetes ? Drugs ? Environmental health ? Infectious diseases ? Maternal health ? Mental health ? Non-communicable diseases ? Obesity and diet ? Occupational health ? Older people ? Physical activity ? Sexual health ? Smoking and tobacco ? Transport ? Vaccine preventable diseases ? Working with and involving communities Treatments, Procedures and Devices Bones and joint surgery ? Cardiovascular surgery ? Cardiovascular system drug treatments ?

Generate Fingerprints

- We use the ‘ssdeep’ fuzzy hash algorithm to generate a fingerprint for the extracted text
  - Compare fingerprints instead of content
- Earlier This Year:
  - aDJjTi6KVkfrehQfnSSXWYjqyBmiF8H9
- From The 2013 Archive:
  - aDJjTi6KWkfrehQfN+SSXWZjbO4kiF+H2LZcn
- Similarity Result: 50%

NICE Example (Archived in 2013)



NHS National Institute for Health and Clinical Excellence

Sign In | Register

Home | News | Get involved | About NICE

Find guidance | NICE Pathways | Quality standards | Into practice | QOF

Home > About NICE guidance > Guidance by type > Clinical guidelines > Lung cancer > CG121 Lung cancer: full guideline appendix 11

NICE guidance

- Find guidance
- NICE Pathways
- Latest guidance
- Guidance by type
- Guidance in development
- How guidance is produced
- NICE guidance research recommendations
- Guidance for patients and public
- Other publications
- Patient safety solutions pilot
- Quality standards

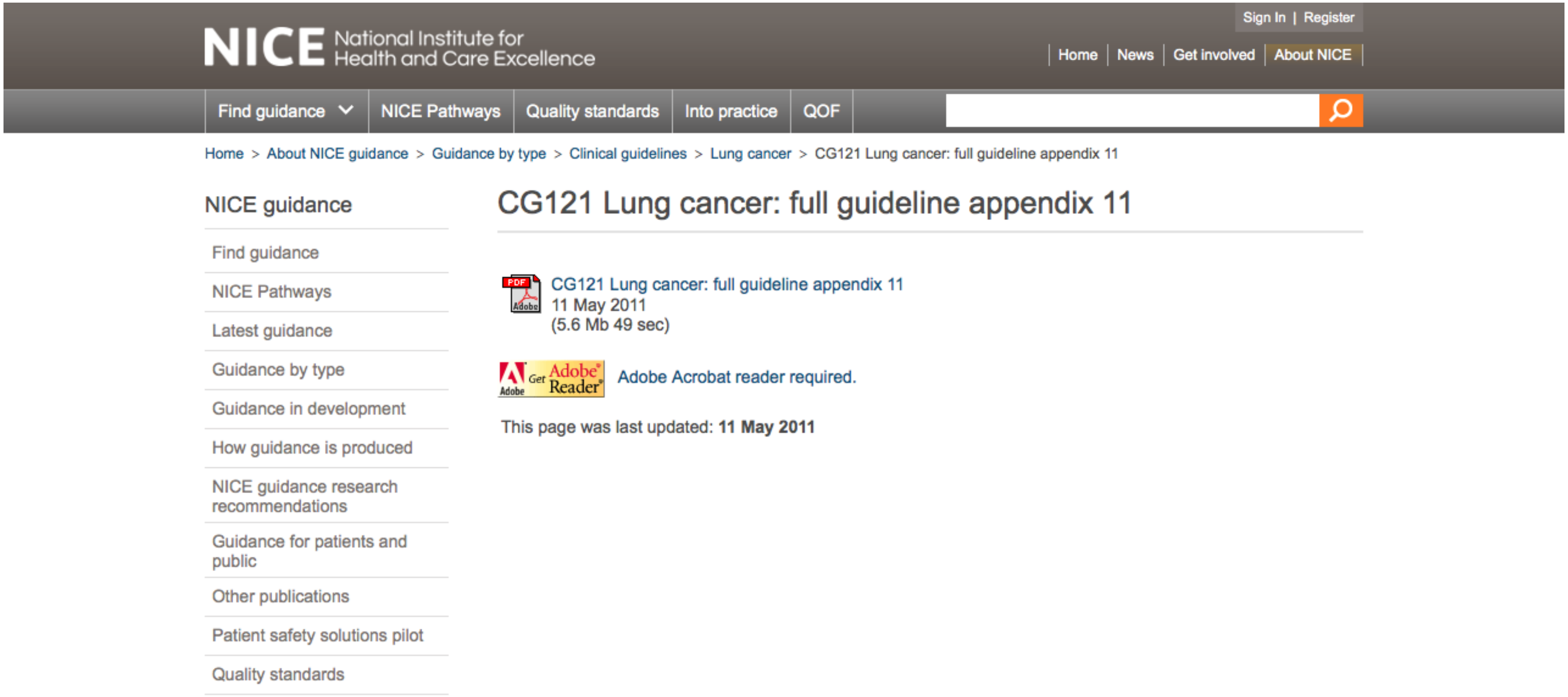
CG121 Lung cancer: full guideline appendix 11

CG121 Lung cancer: full guideline appendix 11
11 May 2011
(5.6 Mb 27 mins 49 sec)

Adobe Acrobat reader required.

This page was last updated: **11 May 2011**

NICE Example (this year)



NICE National Institute for Health and Care Excellence
Sign In | Register
Home | News | Get involved | About NICE
Find guidance | NICE Pathways | Quality standards | Into practice | QOF

Home > About NICE guidance > Guidance by type > Clinical guidelines > Lung cancer > CG121 Lung cancer: full guideline appendix 11

NICE guidance

- Find guidance
- NICE Pathways
- Latest guidance
- Guidance by type
- Guidance in development
- How guidance is produced
- NICE guidance research recommendations
- Guidance for patients and public
- Other publications
- Patient safety solutions pilot
- Quality standards

CG121 Lung cancer: full guideline appendix 11

CG121 Lung cancer: full guideline appendix 11
11 May 2011
(5.6 Mb 49 sec)

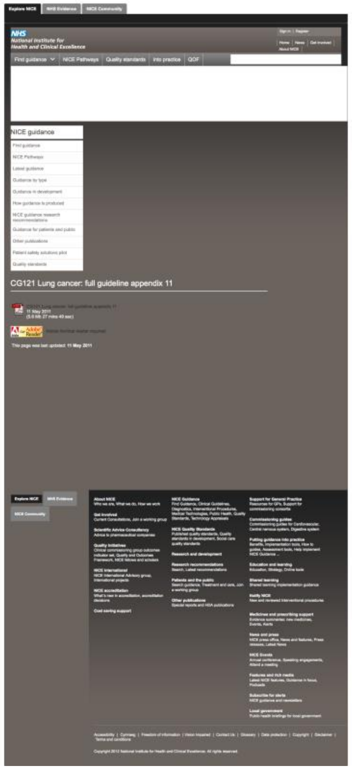
Get Adobe Reader Adobe Acrobat reader required.

This page was last updated: **11 May 2011**

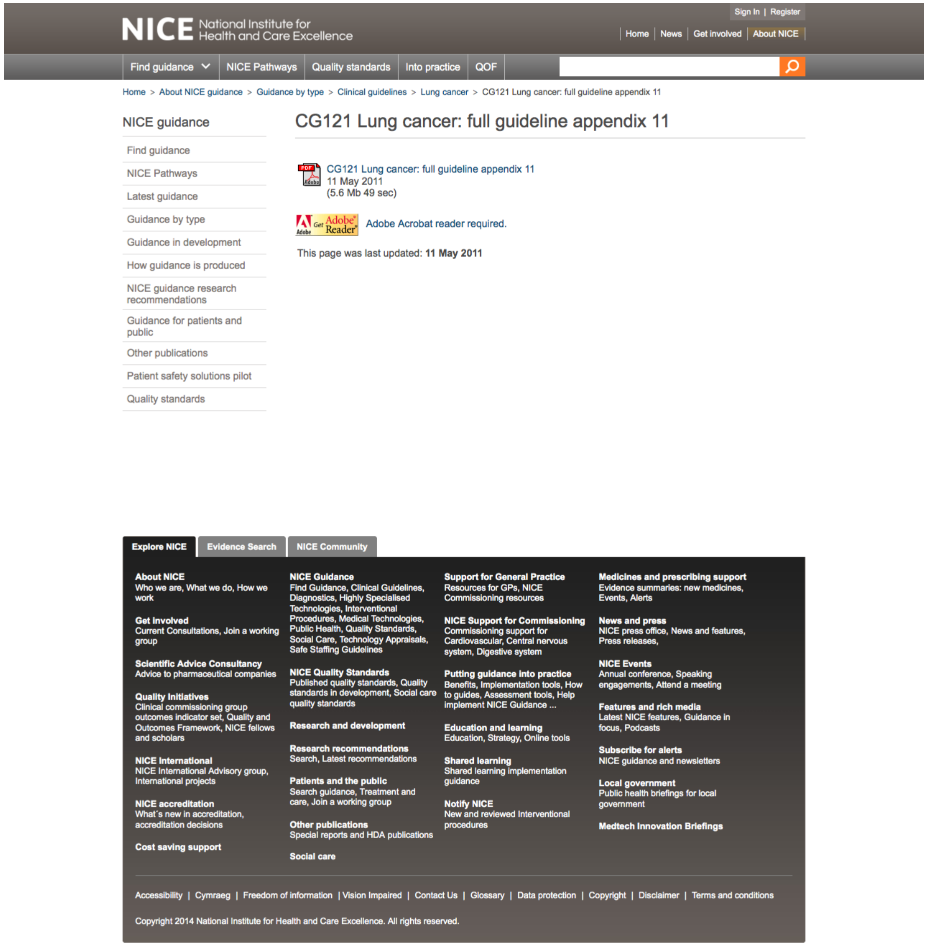
Page Footer Problem

NHS National Institute for Health and Clinical Excellence

CG121 Lung cancer: full guideline appendix 11



CG121 Lung cancer: full guideline appendix 11



Sign In | Register

NICE National Institute for Health and Care Excellence

Home | News | Get involved | About NICE

Find guidance | NICE Pathways | Quality standards | Into practice | QOF

Home > About NICE guidance > Guidance by type > Clinical guidelines > Lung cancer > CG121 Lung cancer: full guideline appendix 11

NICE guidance

- Find guidance
- NICE Pathways
- Latest guidance
- Guidance by type
- Guidance in development
- How guidance is produced
- NICE guidance research recommendations
- Guidance for patients and public
- Other publications
- Patient safety solutions pilot
- Quality standards

CG121 Lung cancer: full guideline appendix 11

CG121 Lung cancer: full guideline appendix 11
11 May 2011
(5.6 Mb 49 sec)

Adobe Acrobat reader required.

This page was last updated: **11 May 2011**

Explore NICE | Evidence Search | NICE Community

About NICE
Who we are, What we do, How we work

Get involved
Current Consultations, Join a working group

Scientific Advice Consultancy
Advice to pharmaceutical companies

Quality Initiatives
Clinical commissioning group outcomes indicator set, Quality and Outcomes Framework, NICE fellows and scholars

NICE International
NICE International Advisory group, International projects

NICE accreditation
What´s new in accreditation, accreditation decisions

Cost saving support

NICE Guidance
Find Guidance, Clinical Guidelines, Diagnostics, Highly Specialised Technologies, Interventional Procedures, Medical Technologies, Public Health, Quality Standards, Social Care, Technology Appraisals, Safe Staffing Guidelines

NICE Quality Standards
Published quality standards, Quality standards in development, Social care quality standards

Research and development

Research recommendations
Search, Latest recommendations

Patients and the public
Search guidance, Treatment and care, Join a working group

Other publications
Special reports and HDA publications

Social care

Support for General Practice
Resources for GPs, NICE Commissioning resources

NICE Support for Commissioning
Commissioning support for Cardiovascular, Central nervous system, Digestive system

Putting guidance into practice
Benefits, Implementation tools, How to guides, Assessment tools, Help implement NICE Guidance ...

Education and learning
Education, Strategy, Online tools

Shared learning
Shared learning implementation guidance

Notify NICE
New and reviewed Interventional procedures

Medicines and prescribing support
Evidence summaries: new medicines, Events, Alerts

News and press
NICE press office, News and features, Press releases,

NICE Events
Annual conference, Speaking engagements, Attend a meeting

Features and rich media
Latest NICE features, Guidance in focus, Podcasts

Subscribe for alerts
NICE guidance and newsletters

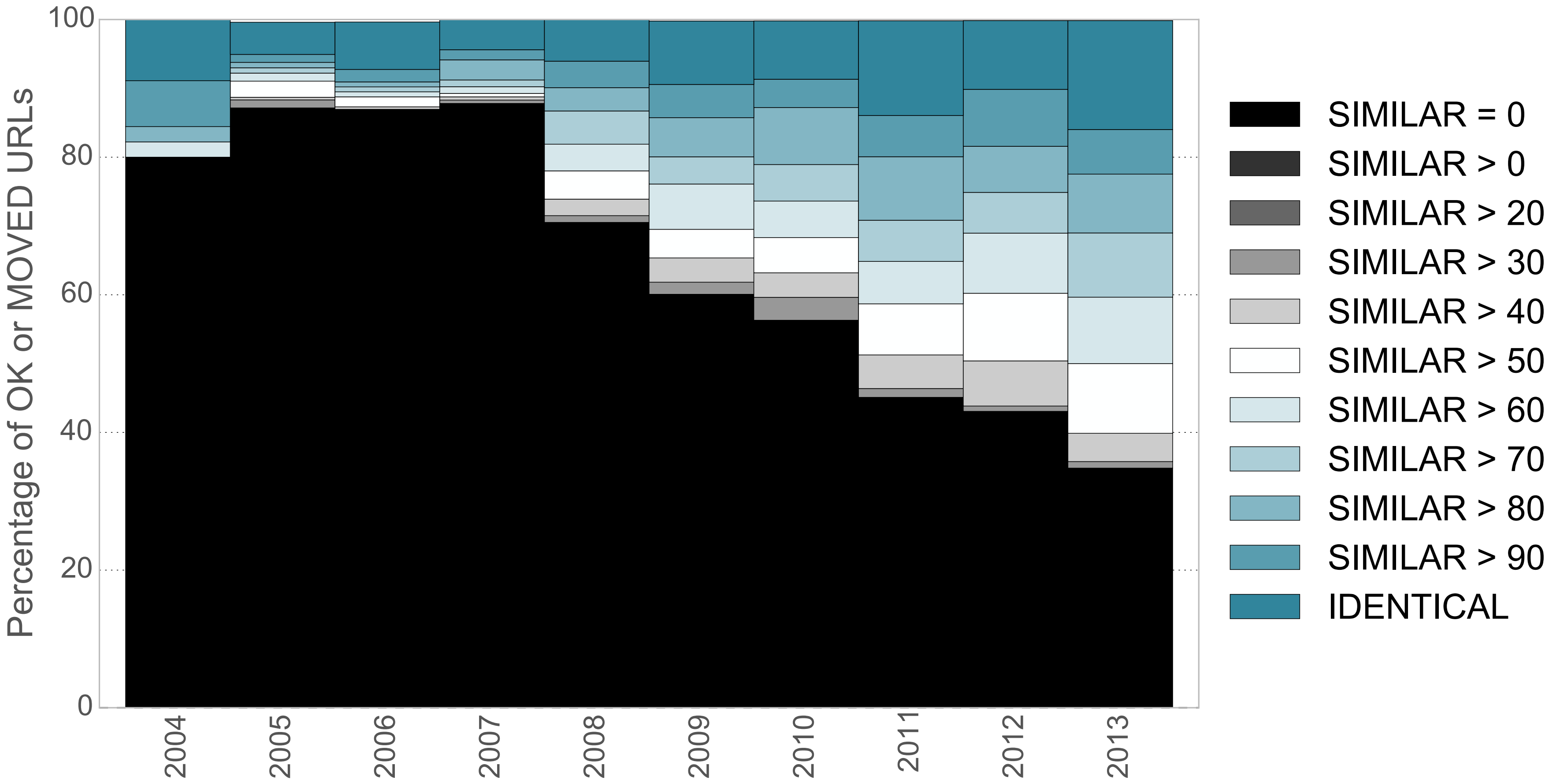
Local government
Public health briefings for local government

Medtech Innovation Briefings

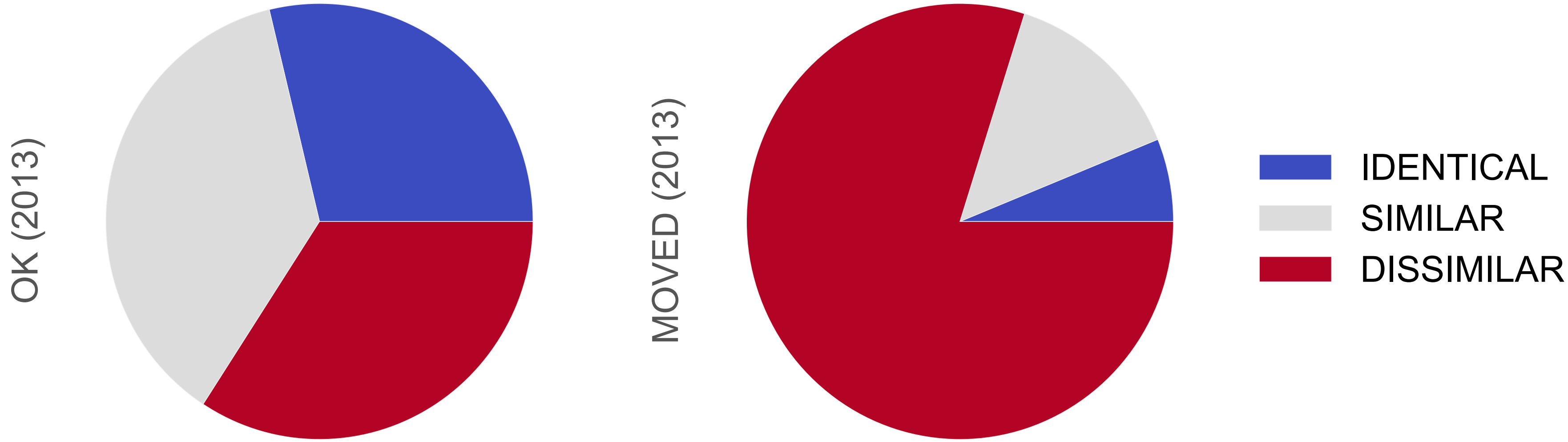
Accessibility | Cymraeg | Freedom of information | Vision Impaired | Contact Us | Glossary | Data protection | Copyright | Disclaimer | Terms and conditions

Copyright 2014 National Institute for Health and Care Excellence. All rights reserved.

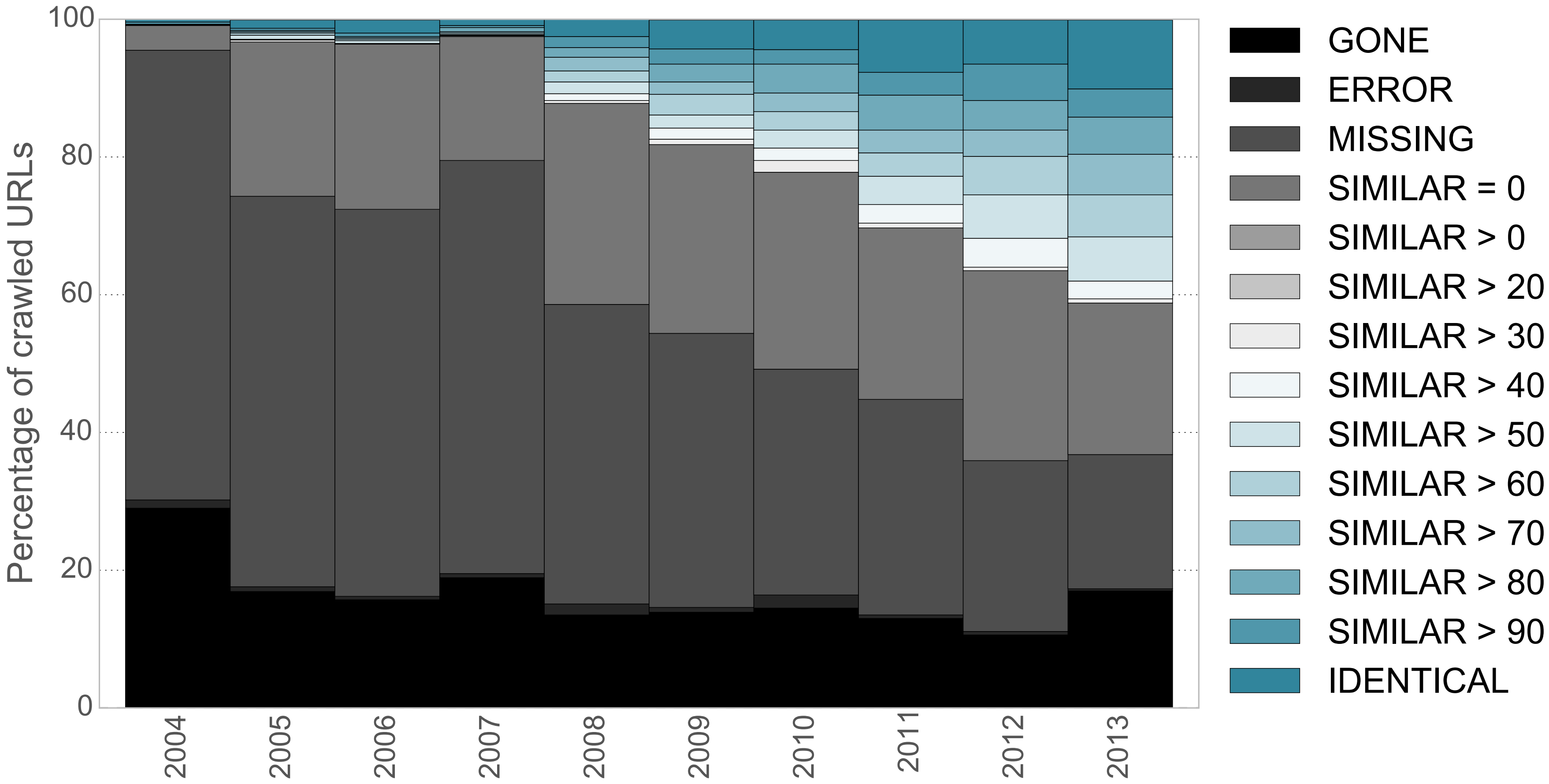
Not Really OK



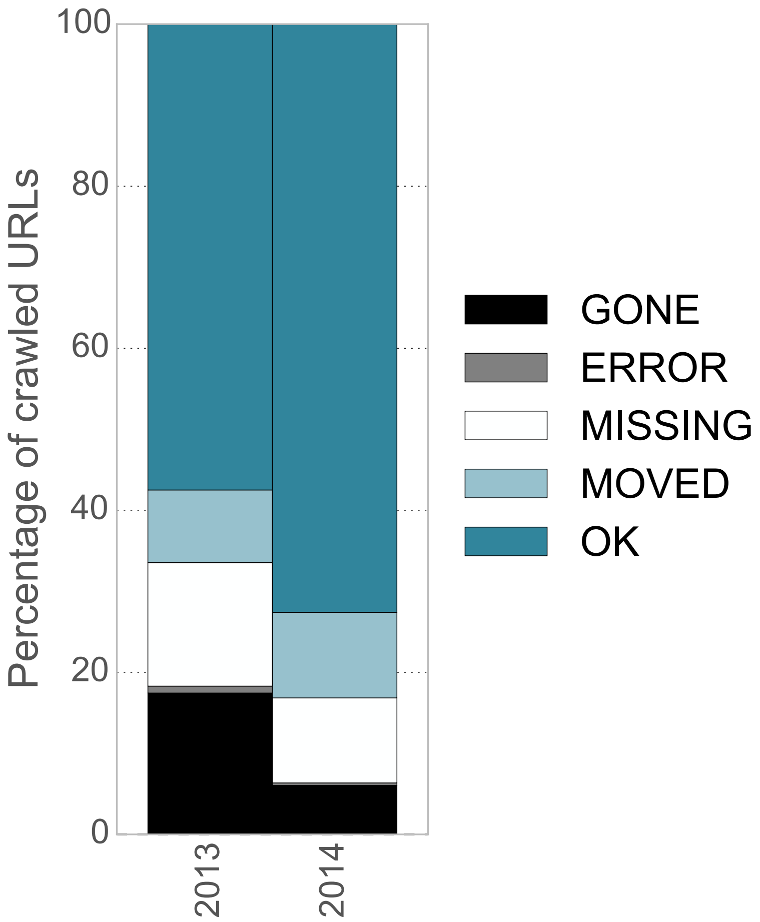
OK versus MOVED

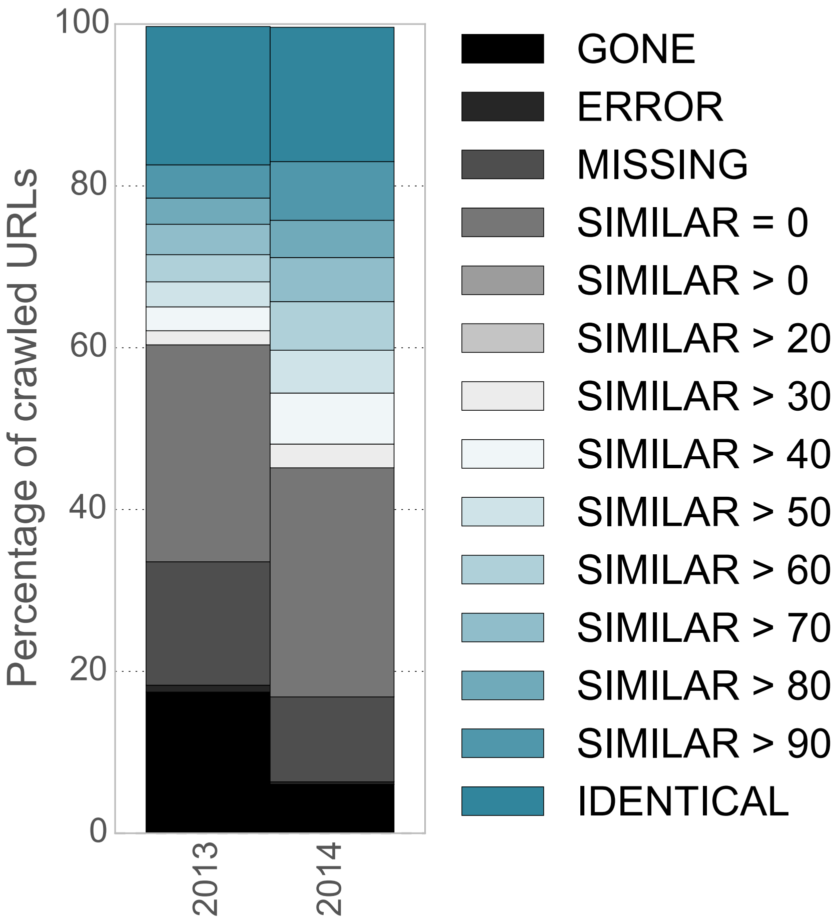


The URLs Ain't Cool

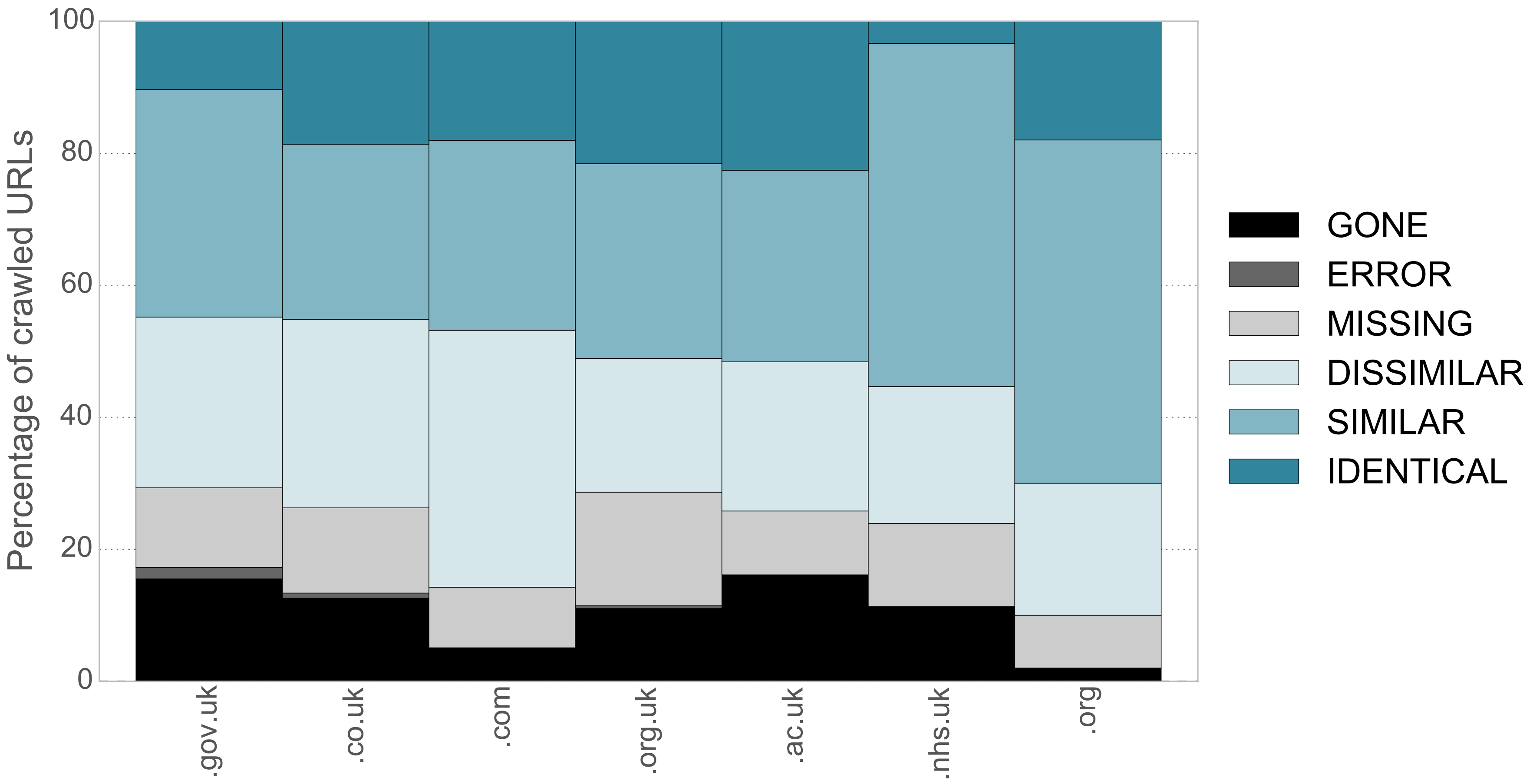


Results For The Legal Deposit Collection

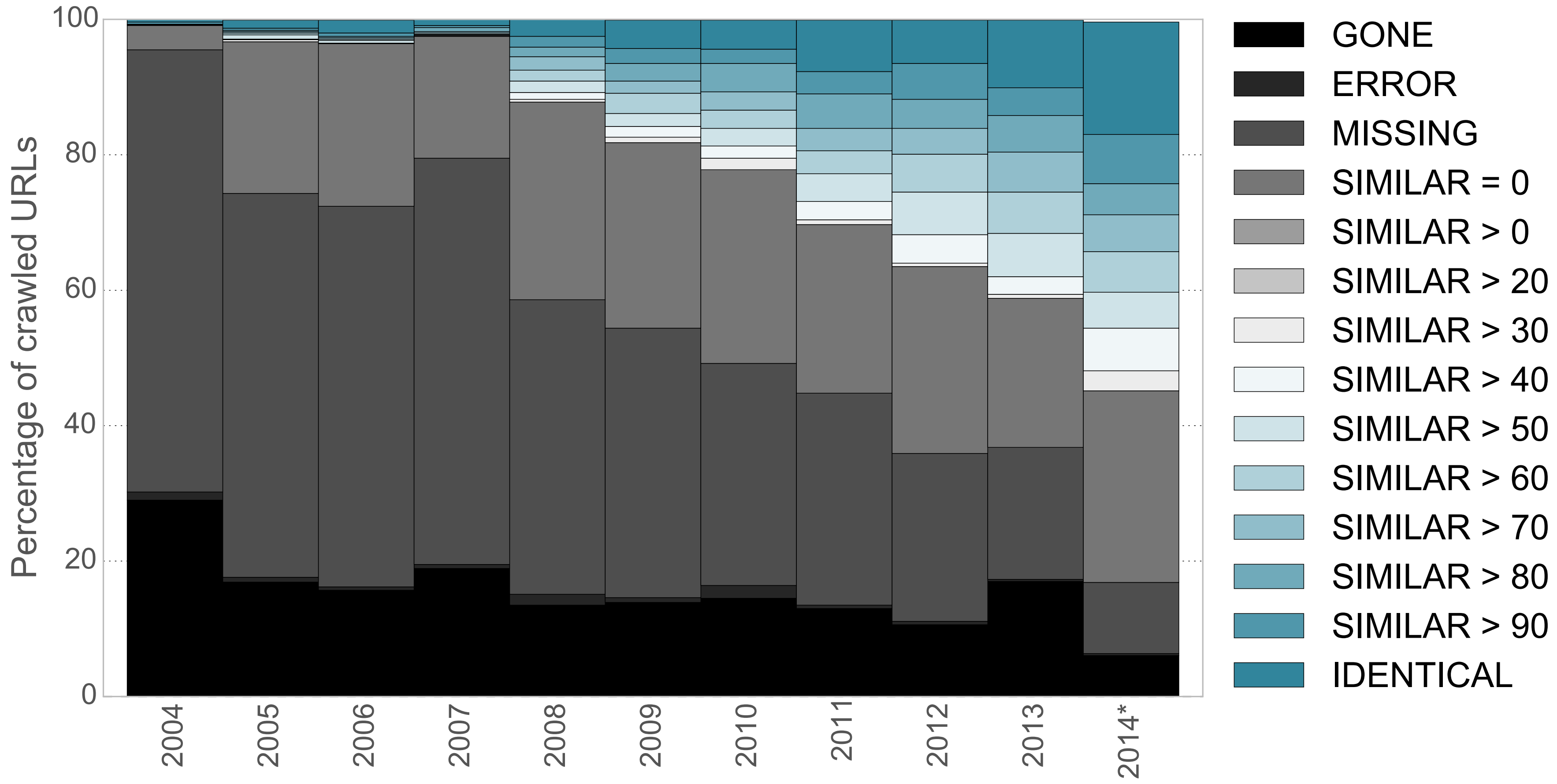




Legal Deposit 2013-2014 By Domain Type



What We’ve Saved (2004-2014)



Summary

- Link rot & content drift dominate:
  - 50% of resources unrecognisable or gone after 1 year
  - 60% after 2 years, 65% after 3 years (islands of stability)
  - Noticeably higher rot rate than results for legal/academic web
- Simple similarity measure provides insight, although:
  - Only sensitive to text changes
  - Overly sensitive to header/footer changes
- Future work:
  - Look for old content at new URLs via hash similarity
  - Compare archival holdings via Memento

Thank you!

Questions?

Getting in touch:

Twitter: @ukwebarchive
Email: web-archivist@bl.uk
UK Web Archive:
http://www.webarchive.org.uk