Digital Preservation and Open Data

Adam Retter

Evolved Binary Ltd

@adamretter / adam.retter@googlemail.com

Adam Retter

- Software Engineer
  - Mostly Scala and Java
  - Anything really...
- Open Source Hacker
- Consultant
- Author
  - "eXist" for O'Reilly
- W3C (Invited Expert)
  - XQuery
  - CSV on the Web, Provenance

Until Recently...

Why Archive?

- Memory is finite
- Stories (facts) become distorted over time
- Point of truth
- Historical Value
- The National Archives
    - Preserving the nations history
    - Governmental Records
    - Public Hearings
    - Special Collections (e.g. Records of LOCOG)

The National Archives

- Archive Records of UK from OGDs, NGOs and Special Interest
- Excellent at traditional Paper records
  - One of the largest collections in the world
  - Over 11 million historical Government and Public Records
- However, most records today are not created on paper!
  - Predicted 2013 - 2020:
    - >6PB of Digital Records to Archive
    - 50% of which will be *Born Digital*
  - 2009: Existing Digital Records System will not cope...
    - 2011: Build new Digital Records Infrastructure = ME :-)

DRI: Digital Repository Infrastructure

1. Records arrive via:
   - Hard Disks (USB etc)
   - DVD / CD / Digital Video Cassette / Tape (mostly LTO 1 to 6)
   - SFTP
2. Load Records
3. Test, Secure and Examine Records (Pre-Ingest)
4. Extract Metadata and Archive (Ingest)
5. Enable Digital Archivists (Search, Retrieval and Edit)
6. Export Transcoded Records and Metadata (Publish / Sell)

Open Source Outputs

- PRONOM - File format database

  http://apps.nationalarchives.gov.uk/pronom

- DROID - File Identification Tool

  http://digital-preservation.github.io/droid

- CSV Schema and CSV Validator

  http://digital-preservation.github.io/csv-schema
  http://digital-preservation.github.io/csv-validator

- UTF-8 Validator

  https://github.com/digital-preservation/utf8-validator

- Shadoop - Scala DSL for Hadoop

  https://github.com/adamretter/shadoop

What is Digital Preservation?

"*In library and archival science, digital preservation is a formal endeavor to ensure that digital information of continuing value remains accessible and usable. It involves planning, resource allocation, and application of preservation methods and technologies, and it combines policies, strategies and actions to ensure access to reformatted and "born-digital" content, regardless of the challenges of media failure and technological change.*"

"***The goal of digital preservation is the accurate rendering of authenticated content over time.***"

- Taken from Wikipedia: https://en.wikipedia.org/wiki/Digital_preservation

What is Digital Preservation?

- File Identification and Analysis / Hardware Analysis
- Emulation vs. Migration
- Multiple copies on diverse media at multiple sites
- Media Retention Policy - Frequently renewed and rewritten
- Pointless without Access?

Why archive Open Data?

- Duh! The same reasons as archiving records.
- Posterity
- Personally: Mining!
  - *As a Digital Archivist, given lots of Open/Linked data over a period of time, I may be able to establish new facts / knowledge*

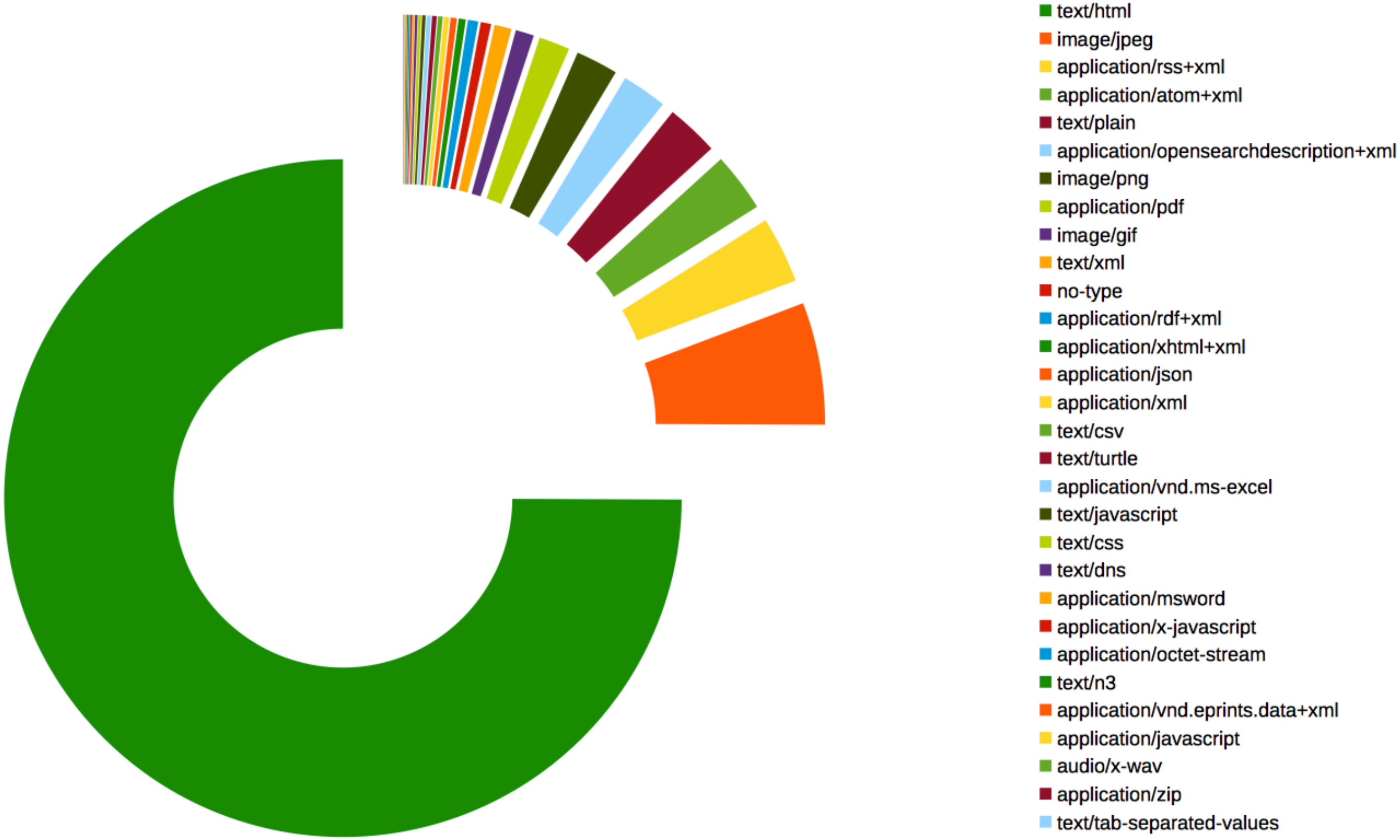
Where to archive Open Data?

- The Internet Archive?

- UK Government Web Archive?
  - http://webarchive.nationalarchives.gov.uk/20140711133430/http://data.gov.uk/

- Private Archive... not very open?

UK Government Web Archive



Problems archiving Open Data

- Web Crawling :-(
  - Web Pages / File Downloads
  - - File Formats e.g. CSV, Excel, PDF, MS Access.
    - Classical Digital Preservation problems!
  - Databases and Query End-points
    - REST
    - SPARQL

- Unstructured Data
  - Context
  - Provenance
  - e.g. CSV Data without headings and/or schema

Linked Data Problems

- Crawling RDF and SPARQL
    - Do Identifiers de-reference?
    - What if links are broken?
    - What if linked dataset is removed/offline
        - Temporary vs Permanent
- Modelling Graph evolution over time

Final thoughts

- Self-describing or described data
  - Some formats are better than others!
  - Human readable Schema?
  - Machine readable Schema?
- Consider provenance
  - Even a timestamp in the data is very useful!
- Is YOUR open data ammenable to crawling?
  - Provide a dump as well maybe?
- How to archive without crawling?