Are research data sets FAIR in the long run?

13th International Digital Curation Conference

20 February 2018

Dennis Wehrle

Klaus Rechert

Albert-Ludwigs-Universität Freiburg

UNI FREIBURG

Agenda

- FAIR principles
- Preserving digital objects / research data
- Research data set analysis
- Sustainability / Preservation risks
- Classification Service
- Lessons learned / Conclusion

FAIR Data Principles

- FAIR principles
  - findable, accessible, interoperable, re-usable
- Problem:
  - Mons et al. (2017):
    - Not a standard
    - Do not specify technical requirements
    - Guiding principles
  - Finding + providing access may not be sufficient for re-usage in the long-run
  - Fast technical life cycle
    - Interoperability, re-use and accessibility depends on data's format + ability to parse / render data

➡ Implies multidimensional problem

Mons, B., Neylon, C., Velterop, J., Dumontier, M., da Silva Santos, L. O. B. & Wilkinson, M. D. (2017). Cloudy, increasingly FAIR; revisiting the FAIR Data guiding principles for the European Open Science Cloud. Information Services & Use(Preprint), 1–8.

Long-term Perspective

- Preserving digital objects
  - Common practice for larger libraries + archives
  - Preservation procedures like file format migrations strategies
- Preserving research data sets
  - Ensuring FAIR in the long run
  - Seems more challenging
    - Higher diversity of file formats
    - Maybe special formats (used by small groups or proprietary data)

➡ “First step” survey:

- Quantify / estimate difficulties by analyzing technical characteristics (file format) of “real life” research data

Data Selection and Preparation

- Repository registry Re3data.org
  - Lists over 1 800 repositories (March 2017)
  - Goal for data selection
    - Randomly select 10 public repositories for the major 14 research disciplines
    - Download approx. 10 data sets for each repository
  - Restriction
    - Only Open Data data sets
    - No prior registration for data access
    - No “frontend only” repository (image galleries, genome database, …)
  - Final selection
    - 92 repositories
    - 3.5 Mio. files (after extracting)
    - 1.95 TB of data

Data Analysis

- Harvard's File Information Tool Set (FITS)
  - Bundles 12 analysis tools
  - Increases detection rate and format coverage
  - FITS analyze one file after another
  - Runtime estimation for a test data set with 9067 files (237 Gb)
    - 19 250 Seconds (~5 hours, 20 minutes)
    - ➡ Over 85 days estimated total runtime
- Reduce runtime
  - 1. Approach: Threads
    - Conclusion: FITS is not thread-safe
  - 2. Approach: Processes
    - Main advantage: Distributed way of working on different (cloud) machines
    - Total runtime: 39 842 Seconds (~11 hours)
      - But not in a single run due to various problems!

Processing Results

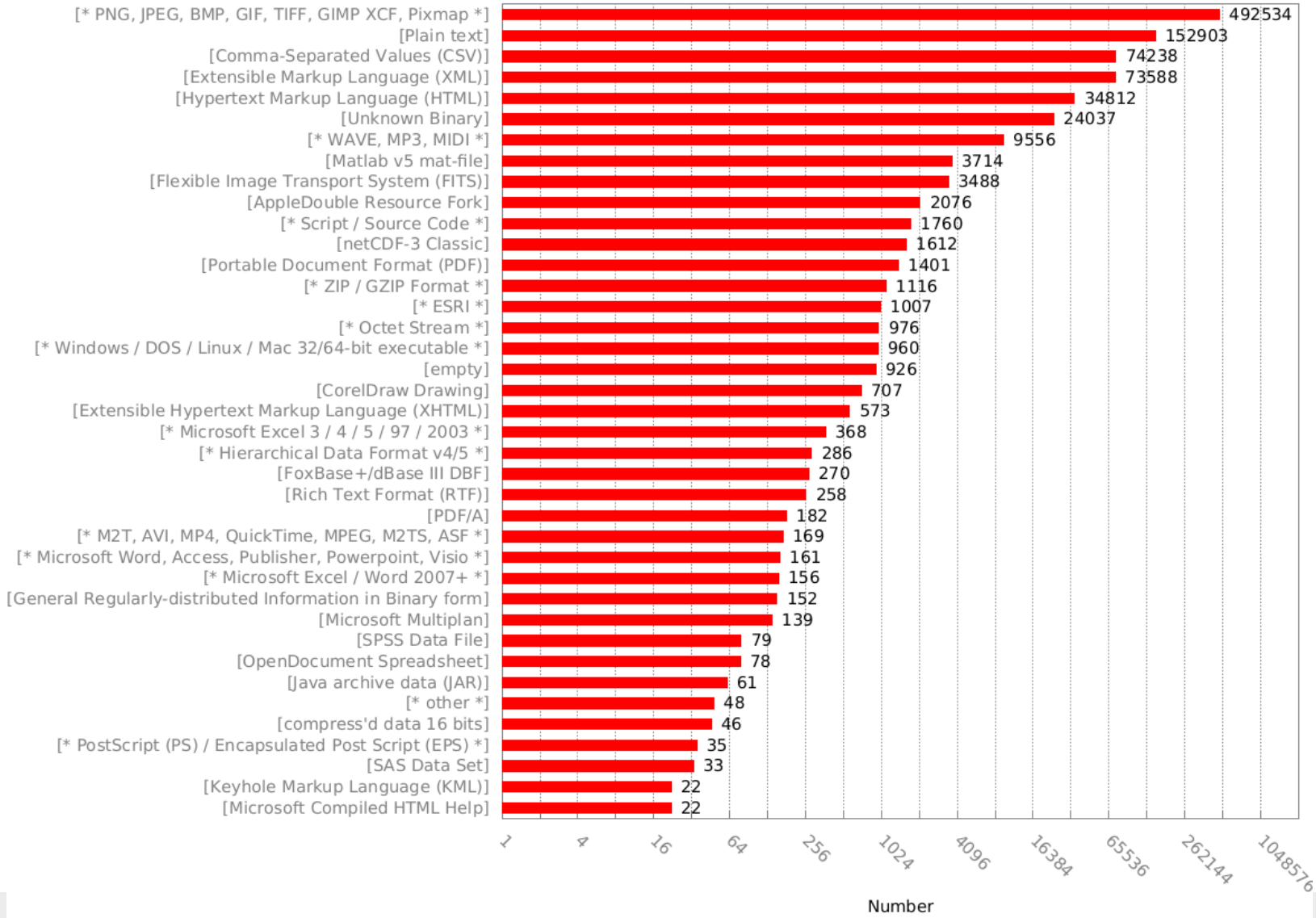
- FITS results
  - No result
  - Single result – tools report same format
  - Conflicting results – tools return different MIME type or format
  - Unknown result – unknown “application/octet stream”
- Post processing (manually)
  - Aggregate identical named formats
  - Unify 260 conflicting results
    - Resolve similar named formats:
      “7-zip archive” vs. “7-zip archive data, version 0.3”
    - Simplifying informations:
      “FoxBase+/dBase III DBF, 2136 records…” ➡
      “FoxBase+/dBase III DBF”
    - Resolve differently named formats:
      “Netpbm image data, bitmap” vs. “Portable Bit Map” ➡
      “Portable Bitmap”

Processing Results

- Post processing (cont.)
  - Not resolved: 28 conflicts (2 150 files)
    - E.g.: “Plain text”,”M2T” => no M2T video files, but rather SPS data files
    - Excluded those 2 150 files from result, mostly due to lack of domain specific knowledge
    - Example of problematic data sets
      - Even if FAIR now, what about in long-run?
- Overall 145 formats identified (lower estimation)

Result*

* ~1.6 mio. XHTML and 957809 XML files excluded (reason: two big data sets, which strongly would distort this figure)

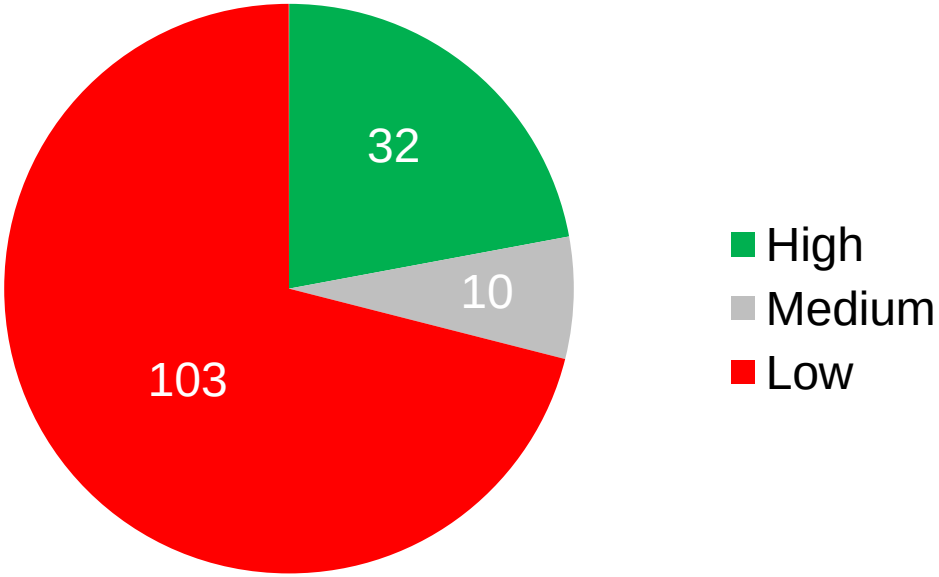


Discussion

- Images (PNG(437 855), JPEG (46 679), …)
- Text-encoded formats like CSV, XML, RTF, HTML and script/source code
  - Usually readable with simple text editor
  - BUT: interpretation not guaranteed
    - Base64-encoded binary data in XML
    - JavaScript in HTML
    - Reference to external data in (X)HTML
- More problematic “text files”
  - Source code requires build- or runtime environment
  - Complete system environment: Matlab, SPSS Data, Octet Stream, unknown Binary, 32/64-bit executable
- Quantify sustainability / preservation risks?

Sustainability / Preservation Risks

- Cornell University[1] – divide file formats into
  - High probability (plain text, PDF/A, PNG, …)
  - Medium probability (OpenOffice, GIF, …)
  - Low probability (WordPerfect, Word (.doc), …)
- Applied to data set



[1] Recommended File Formats, http://guides.library.cornell.edu/ecommons/formats

Classification Service

- bwDATA Diss
  - Preserve dissertation data with dissertation text
  - Assess preservation risks of data sets
  - Characterization service[1]
    - Simple traffic-light visualization
    - Based on a individual policy file
  - Characterization result
    - Pre-ingest check
    - Advise to re-consider file format choice (if possible)
    - Raise awareness on un-sustainability of format
    - Guide software collection to render data sets
    - Prepare emulation / virtualization strategy


```
{
    "value": "GREEN",
    "type": "fmt/44",
    "count": 261,
    "typeName": "JPEG File Interchange Format",
    "fromDate": 845919420000,
    "toDate": 845919420000
}
```

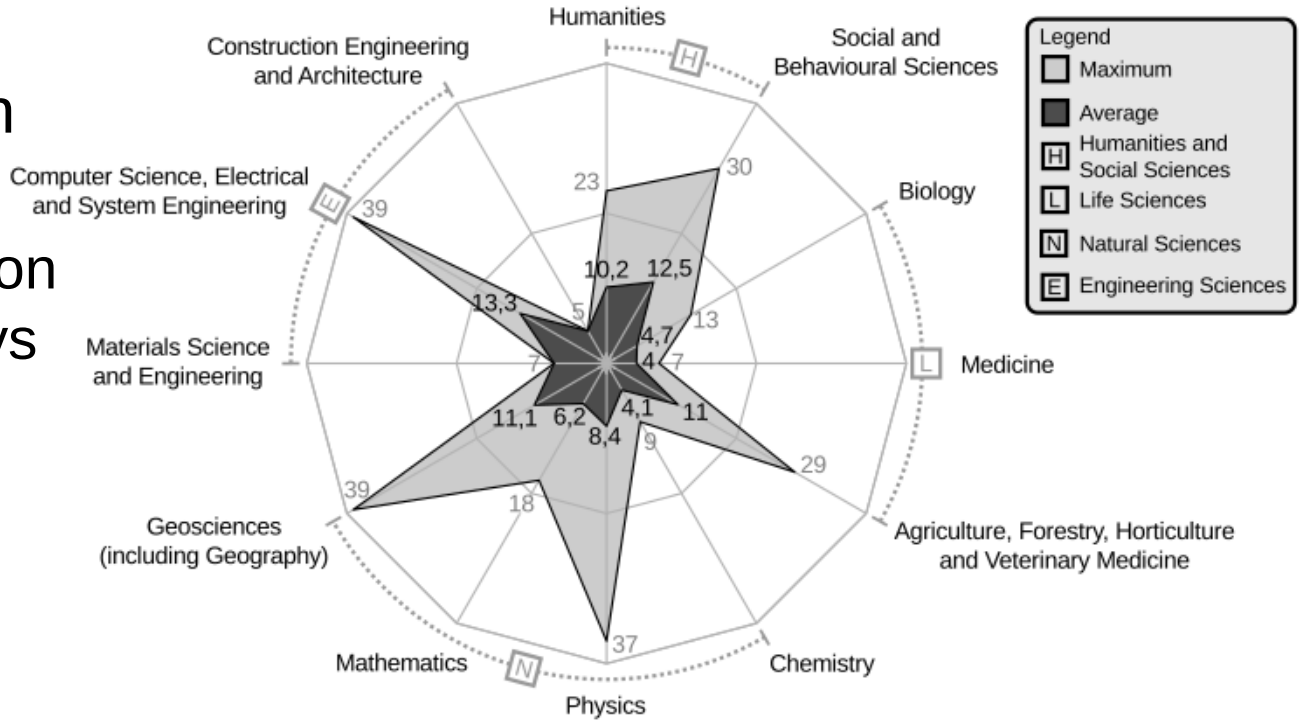
```
{
    "value": "YELLOW",
    "type": "fmt/39",
    "count": 1,
    "typeName": "Microsoft Word Document",
    "fromDate": 845919420000,
    "toDate": 845919420000
}
```

```
{
    "value": "RED",
    "type": "fmt/564",
    "count": 1,
    "typeName": "Adobe Illustrator",
    "fromDate": 845919420000,
    "toDate": 845919420000
}
```


[1] Classification Service, http://classifier.eaas.uni-freiburg.de

Data Centric View to Data Processing View

- Classic migration-driven approach
  - focuses on individual file formats
- Research data sets are more heterogeneous
  - Re-usage: File formats may have strong interdependencies
- Preservation planning / action
  - Single file format migration may not always be sufficient



Lessons learned / Conclusion

- FAIR
  - Important concept for long-term preservation
  - Even though it is abstract, it fosters to think about it
  - Multidimensional problem ➡ no single solution
- “simple” file format analysis
  - Harder than anticipated
  - Tool support was weak
  - Handling large amount of data is challenging
  - Manually steps were necessary
- File Formats
  - Generic research data service can’t refuse badly rated formats
  - Creator should be involved