IJDC | *General Article*

Are Research Datasets FAIR in the Long Run?

Dennis Wehrle
University of Freiburg

Klaus Rechert
University of Freiburg

Abstract

Currently, initiatives in Germany are developing infrastructure to accept and preserve dissertation data together with the dissertation texts (on state level – bwDATA Diss[1], on federal level – eDissPlus[2]). In contrast to specialized data repositories, these services will accept data from all kind of research disciplines. To ensure FAIR data principles (Wilkinson et al., 2016), preservation plans are required, because ensuring accessibility, interoperability and re-usability even for a minimum ten year data redemption period can become a major challenge. Both for longevity and re-usability, file formats matter. In order to ensure access to data, the data's encoding, i.e. their technical and structural representation in form of file formats, needs to be understood. Hence, due to a fast technical lifecycle, interoperability, re-use and in some cases even accessibility depends on the data's format and our future ability to parse or render these.

This leads to several practical questions regarding quality assurance, potential access options and necessary future preservation steps. In this paper, we analyze datasets from public repositories and apply a file format based long-term preservation risk model to support workflows and services for non-domain specific data repositories.

1 BwDATADiss-bw Data for Dissertations: https://www.alwr-bw.de/kooperationen/bwdatadiss/
2 EDissPlusDFG-Project – Electronic Dissertations Plus: https://www2.hu-berlin.de/edissplus/

Received 22 January 2018 ~ *Accepted* 22 January 2018

Correspondence should be addressed to Dennis Wehrle, Hermann-Herder-Str. 10, 79104 Freiburg. Email: dennis.wehrle@rz.uni-freiburg.de

An earlier version of this paper was presented at the 13th Digital Curation Conference.

The *International Journal of Digital Curation* is an international journal committed to scholarly excellence and dedicated to the advancement of digital curation across a wide range of sectors. The IJDC is published by the University of Edinburgh on behalf of the Digital Curation Centre. ISSN: 1746-8256. URL: http://www.ijdc.net/

Introduction

According to the FAIR data principles, data should be findable, accessible, interoperable and re-usable (Wilkinson et al., 2016). These four basic principles are a vital requirement to enable and foster re-use of research data and should form the base for validation of research results as well as to formulate new and maybe interdisciplinary research questions.

The FAIR principles are not fully specified and remain open for interpretation (Mons et al., 2017), in particular when implementing research data services. FAIR refers "to a set of principles, focused on ensuring that research objects are reusable, and actually will be reused, and so become as valuable as is possible. They deliberately do not specify technical requirements, but are a set of guiding principles that provide for a continuum of increasing reusability, via many different implementations" (Mons et al., 2017). Finding and providing access to research data is an important requirement, but simply searching and downloading research data may not always be sufficient to re-use them in the long run. Scientist already have to invest a significant amount of their time for data preparation (cleaning, organizing and collecting data) (Press, 2016). Future researchers might need to invest even more effort to re-use research data, by decoding obsolete file formats. Hence, due to a fast technical lifecycle, interoperability, re-use and in some cases even accessibility depends on the data's format and our future ability to parse or render such data formats.

Preserving digital objects is now common practice for larger libraries and archives. They have implemented preservation procedures such as file format migrations strategies for their digital collections. Preserving research datasets seems, however, more challenging, as one can assume a much higher diversity of file formats compared to collections found in libraries. Furthermore, there might be a variety of special formats, only used by small user groups or proprietary data emitted from special purpose machinery.

To be able to quantify or estimate the difficulties of preserving research data, we have analyzed the technical characteristics (file format) of 'real life' research data found in public repositories. As result of this analysis we have developed a simple traffic-light based format risk assessment service for research datasets, to provide feedback to researchers when submitting datasets and to reflect preservation risks of already submitted datasets.

Research Data Diversity

Recently, studies and surveys on various aspects of research data management practice have been published.

Kennan and Markauskaite (2015) conducted a large study on data management practice by addressing academics directly. Also, recent studies like Tristram et al. (2015) or Simukovic, Kindling and Schirmbacher (2013) focus on academics directly. For instance, Paul-Stüve, Rasch and Lorenz interviewed 218 members of Kiel University about file formats contained in their datasets. 52.29 % use discipline- or device-specific data, 50.46 % spreadsheets, 47.25 % text documents, 46.33 % databases, 36.70 % images and 27.52 % programs and applications.

Austin et al. (2015) surveyed online services for storage, curation and sharing for research datasets. Most services focus is on access and sharing (collaboration) while long-term accessibility is usually ensured through bit-preservation services and some format-specific file format migration services. Furthermore, publication and sharing of scientific workflows (Atkinson, Gesing, Montagnat and Taylor, 2017) and in particular reproducible research has gained momentum (Peng, 2011), but the long-term perspective of these concepts and tools require even more attention.

Woods and Brown (2008) analyzed file formats regarding file format migration options of a large CD-ROM collection. Woods and Brown identified format migration paths for roughly 25% of the total files in the dataset, while only 33% of the files did not require migration (ASCII and HTML formats). While the identified migration paths could be used with a high success rate, still a quite large proportion of files (and file formats) could not be migrated, either due to complex and proprietary data formats (e.g. older Office formats) and (unknown) binary files.

Roche, Kruuk, Lanfear and Binning (2015) analyzed "100 data sets associated with nonmolecular studies in journals that commonly publish ecological and evolutionary research and have a strong public data archiving policy". Roche et al. conclude that "out of these data sets, 56 % were incomplete, and 64 % were archived in a way that partially or entirely prevented reuse".

Because of the interdisciplinary nature of collected research data and due to the lack of domain specific knowledge of generic data repositories, we don't measure the quality of research data (on content level), but rather focus on the technical characteristics of research datasets and their (successful) preservation probability.

Data Selection and Preparation

To investigate the file format breadth and diversity of research data sets we have used the re3data registry[3] with the intention to analyze different repositories for every main research discipline[4] (see Figure 3 for a list of selected disciplines). From over 1,800 listed repositories (time of analysis on March 2017), our intention was to randomly select ten repositories for each discipline and download approximately ten datasets from each repository. For practical and legal reasons, the selection was restricted to Open Data datasets and repositories, which did not require a prior registration to access data. Furthermore, repositories that only provide a frontend to access a database, e.g. to display data in the browser or to produce image galleries, were ignored.

Our final selection consisted of 92 repositories, since the intended number of ten different repositories could not be met for each discipline. For instance, no suitable repositories could be identified for Mechanical and Industrial Engineering and Thermal Engineering. On the other hand, there were many potential repositories for Medicine, but most repositories restricted access to data.

After downloading the research datasets, as an initial preparation step, we recursively extracted all archives like *.tar.gz, *.zip, etc. and deleted operating system specific folders like DS_Store and __MACOSX. Altogether our final sample consisted of 3,509,511 individual files resulting in 1.95 Tb of data. File size of individual files ranged from 0 (we have found 926 empty files) to a single 32.36 Gb file. Nearly 14,000 files had a file size of 283 Bytes and the average file size was 555 Kb. Figure 1 shows

3 Registry of research data repositories: https://www.re3data.org/

4 Browse re3data.org by subject at: https://www.re3data.org/browse/by-subject/

the distribution of the individual file size (note, that because of logarithmic scale, we added file size of 0 manually).

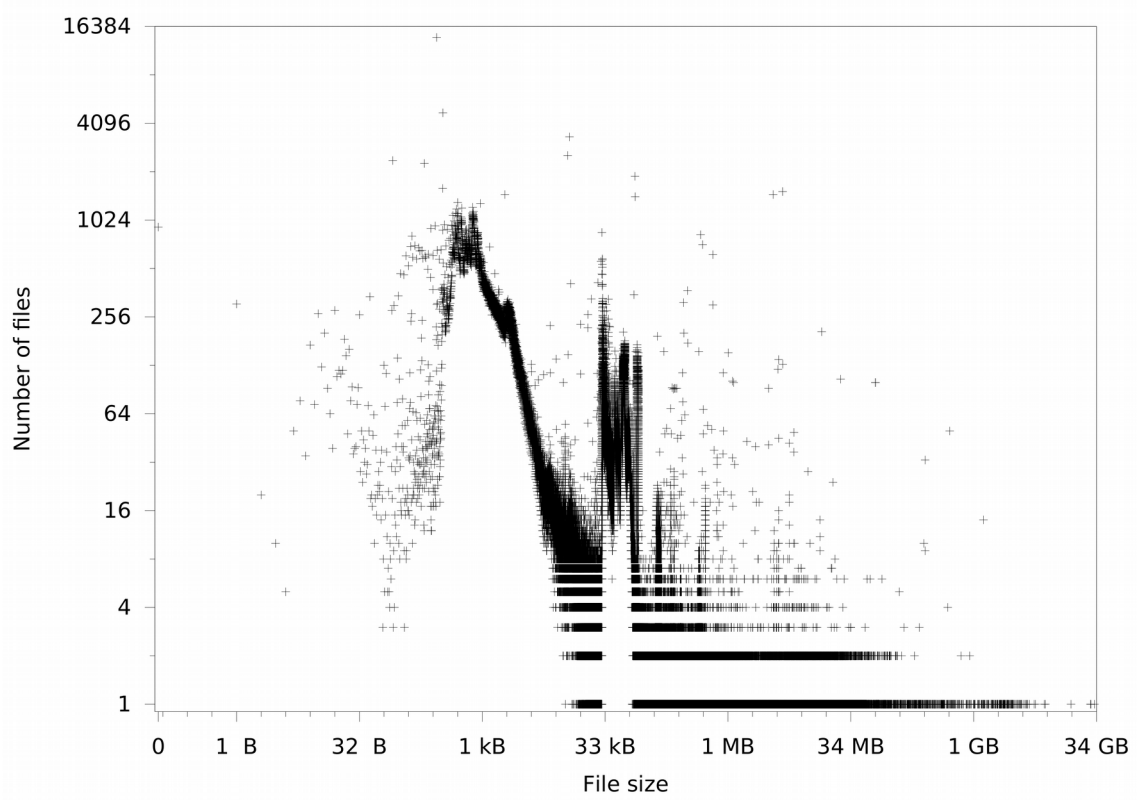

Figure 1. Distribution of file sizes with logarithmic scale on x- and y-axes.

In order to determine technical information, we have chosen Harvard's File Information Tool Set (FITS)[5], as FITS bundles different analysis tools (currently 12) and thus increases detection rate and format coverage. By default FITS uses 20 internal threads to analyze one file with different tools in parallel. But because only one file is analyzed at the time, i.e. free threads won't start analyzing another file, the runtime for analyzing one file is bound by the slowest tool. Hence, the total time needed to analyze a dataset is the sum of time needed to subsequently analyze one file after another. To estimate the total runtime, we chose a random data set from Computer Science, Electrical and System Engineering containing 9,067 files (237 Gb). The analysis of the data set took 19 250 seconds (5 hours; 20 minutes; 50 seconds) with an average analysis time of 2.21 seconds per file. Based on this number, the analysis of all downloaded data sets would take at least 85 days.

To reduce the required time, we decided to speed up the analysis by using threads in order to start multiple FITS instances to analyze multiple datasets in parallel. However, no stable (reproducible) FITS run on a dataset was possible. A closer investigation showed that FITS is not thread-safe. For instance, the MediaInfo tool uses a static file handle and thus all threads share the same file handle, preventing multiple threads analyzing multiple files. Since FITS includes multiple external libraries and tools, making FITS thread-safe is non-trivial and was not considered for this work.

5 File Information Tool Set (FITS): https://projects.iq.harvard.edu/fits/home

In order to increase parallelism, we chose a different approach and wrapped FITS into individual processes and developed a governor process orchestrating the FITS worker processes as well as collecting their results. In contrast to threads, which share resources, processes are strictly separated. This allows to run several FITS instances simultaneously. Technically this comes with a price. Within threads, one thread is able to access the data from another thread. Thus the main thread was able to collect the result and put it into a special data structure. To be able to aggregate the result from one dataset in our processes driven approach, we had to develop a so-called cluster architecture. Within this kind of architecture, there exists one governor instance and several worker instances. The main task of the governor instance is to assign analyzing jobs to workers as well as supervising worker instances and process their results. Every worker receives up to ten analyzing jobs and sends the result to the governor instance. Thus, we could successfully run FITS processes in parallel on cloud machines. The final bottleneck then was getting the data fast enough to the analyzing instances. A lot of the analysis time on cloud machines (VM1 and VM2) was wasted by waiting for data, which used a networked (NFS) storage backend. To actually cope with such an amount of data, we also included a specialized (dedicated) hardware.

Table 1. Configuration of hardware used for analyzing research data sets.

	Virtual Machines VM1[a] and VM2[b]	Dedicated Server
CPU	8[a] / 4[b] Cores (Intel Xeon E5-2640 v3 @ 2.60 GHz)	40 Cores (2 x 10 Core Intel Xeon E5-2630 v4 @ 2.20 GHz)
RAM	16 Gb[a] / 8 Gb[b]	256 Gb
Data Access	Network (NFS) @ 10 Gbit/s	SSD RAID (20 x Intel 53210 1.6 Tb)
OS	Ubuntu 16.04 LTS (4.4.0 Kernel)	

Based on technical configuration (cf. Table 1) the dedicated server was configured to run the governor instance as well as 40 worker instances. Additionally, VM1 ran sixteen and VM2 eight worker instances. Using this setup the aforementioned test data set with 9,067 files was processed in 684 seconds (average 0.075 seconds per file). Compared to the usage of a standard Desktop-PC using a more powerful dedicated server reduced the runtime marginally by 7.25 % (cf. Table 2). By analyzing the same data set in parallel with 64 instances, we were able to reduce the runtime by 96% (compared to single-threaded approach on dedicated server). Processing all chosen data set took 39,842 sec (11 hours, 4 minutes, 2 seconds) instead of the initially estimated 85 days.

However, the stated total time has never been achieved in a single run on all datasets. Various software problems and limitations of the analysis tools prevented a complete run. For instance, dataset #6 of the Computer Science, Electrical and System Engineering sample contained 1.6 mio XHTML files and no other formats. A JHOVE validation constantly took exactly 1:01 minutes per file, indicating a bug or timeout loading external resources. A further problem with JHOVE was memory consumption in certain datasets. At least 4 Gb RAM per instance were necessary to successfully analyze all datasets with JHOVE enabled. Furthermore, we experienced long-running processes of the ExifTool analyzing XML files, even though this tool's purpose is to

extract metadata from image or video files, e.g. we found the ExifTool analyzing a 16 Mb XML file for several hours.

By default, FITS returns for each file 'no result', 'single result', 'conflicting results' or 'unknown result'. Conflicts may appear if different tools return different MIME types or formats for a single file. Unknown results are typically MIME types 'application/octet- stream' with no additional file format information. In contrast to the FITS single XML output per file, we stored the result of a complete dataset into a manually editable CSV file. In order to build the complete result, we had to manually aggregate the result from each single dataset into a consolidated result file. After all results have been collected, as a first step, identical named formats were aggregated. In a second data aggregation step, conflicting results have been unified in several steps. This included to resolve similar named formats like [7-zip archive] vs. [7-zip archive data, version 0.3] as well as simplifying informations like [FoxBase+/dBase III DBF, 2136 records ...] into [FoxBase+/dBase II DBF]. As a final step, we have resolved conflicts where file formats were named differently, such as [Netpbm image data, bitmap] and [Portable Bit Map] to [Portable Bitmap]. In our sample, we have found more than 260 such conflicts concerning 8,282 files. After manually resolving this conflicts, we still had 28 conflicts affecting 2,150 files.

Table 2. Analysis runtime for test data set (9 067 files). Runtime reduction from Dedicated Server relating to Desktop-PC, from Cluster relating to Dedicated Server. Cluster = VM1 (16 instances), VM2 (8) and Dedicated Server (40) with a total number of 64 instances.

	Single-Thread		Cluster
Runtime	Desktop-PC	Dedicated Server	64 Instances
Total	19,250 sec	17,854 sec	684 sec
Per File	2.12 sec	1.97 sec	0.075 sec
Reduction		1,396 sec (7.25%)	17,170 sec (96.45%)

Manually reviewing some conflicts showed that the conflict [[Plain text],[M2T]] doesn't contain M2T video files but rather SPS data files which were wrongly identified by ExifTool. Also, [[Plain text],[* Portable Pixmap, Graymap, Bitmap *]] weren't images at all but text files. Since not all conflicts could be resolved (mostly due to a lack of domain specific knowledge), we have excluded these files from further analysis steps.

File Format Analysis and Discussion

After post-processing, more than 140 distinct file formats have been identified. In order to create a more compact view on data formats found, we have grouped similar file formats, e.g. the group image formats consist of PNG, JPEG, BMP, GIF, TIFF, GIMP XCF and Portable Pixmap. Furthermore, we have created a special group [* other *], which contains file formats which were found only in single digit numbers (e.g. TrueType Font (8), JavaScript (5), FPX (4), AutoCAD (4), Adobe Photoshop (3), SVG (3), WordPerfect (3), etc.). Even though we have used a wide variety of tools (FITS included 12 different tools) 24,037 files still remain unknown. Figure 2 shows the

distribution of file formats found in the research datasets. Note that we excluded two big data sets from Figure 2 containing 1,668,341 XHTML and 957,809 XML files, because these two datasets would strongly distort Figure 2 as well as diminish the importance of the other formats by indicating, that these formats can be neglected. The two big datasets in question consisted of web crawls representing both Computer Science and Social Science.

In our sample, after excluding the aforementioned exceptional datasets, the images format group (PNG (437,855), JPEG (46,679), etc) is followed by a number of different text-encoded formats including CSV, XML, RTF, HTML and script/source code. In general plain text-based formats are usually readable with a simple text editor (we neglect character encoding issues here), hence, access to the information content should be partially ensured. However, even with simple formats like these, in general it can not be guaranteed that the information can be interpreted correctly. For instance, XML files could contain base64-encoded binary data, HTML files may contain JavaScript elements, which requires a suitable runtime (web browser) to display information or functionality and furthermore, (X)HTML files may include external references to data or other content.

While text-based formats are system- or platform-independent and usually can be viewed or interpreted with a variety of programs, some subgroups of text-based formats are more problematic. For instance, source code requires a build- and/or a runtime-environment. Even though one can extract specific information from source code by using a simple text editor, e.g. parameters or settings used for a specific algorithm, (re-)building or compiling the code to an executable requires additional software. Similar, file formats such as [Matlab v5 mat-file] or [SPSS Data File] require additional software for interpretation. This also applies for the unknown [Octet Stream] (976) and [Unknown Binary] (24 037) files. Additionally, the [* Windows / DOS / Linux / Mac 32/64-bit executable *] (960) files require a complete system environment, as they are platform specific executables. Also old formats like [Microsoft multiplan] illustrates this requirement.[6] Hence without a specific environment, it is not possible to re-use such data.

Additionally we found compressed formats like [ZIP / GZIP Format] (1,116) which weren't files with a *.zip extension but formats which used some kind of compression. Random inspection showed that, among others, Google Earth KMZ files were identified as ZIP files because they are compressed Keyhole Markup Language (KML) files (with *.kmz extension). From 128 GZIP files, 47 files are in RData[7] format. The remaining 81 files are indeed files with a *.gzip extension, but all of them were corrupt and couldn't be extracted. In total we found four corrupted compress'd files and 81 corrupted GZIP archives, as well as 926 empty files.

6 Multiplan is an old spreadsheet program, which was originally designed for DOS in 1982 and was later released for Apple II and Comodore 64.

7 RData is an old file format of the statistical software R.

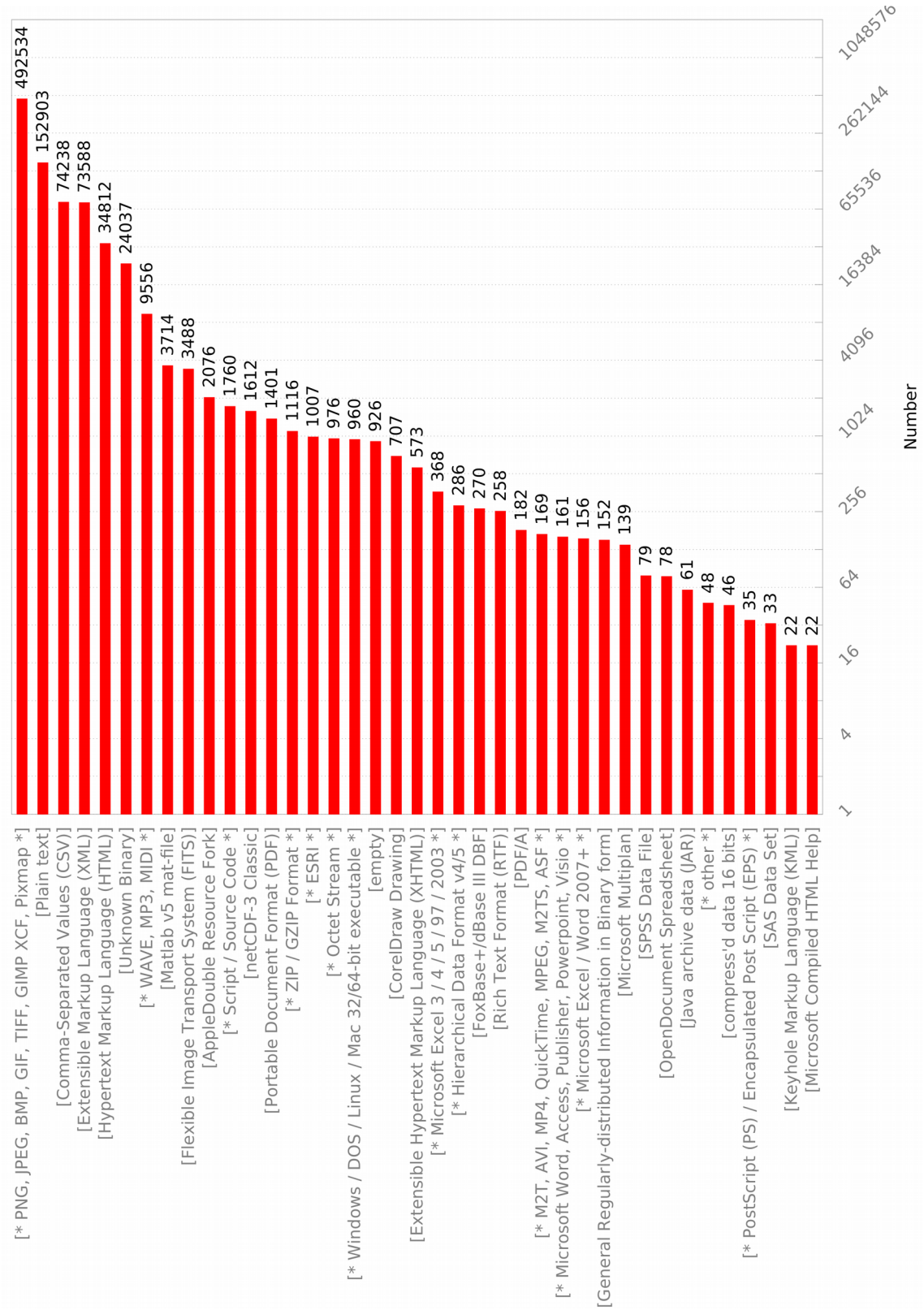

Figure 2. File format distribution. Grouped file formats are denoted with [*...*].

In order to quantify the preservation problems for our data sample, we assessed sustainability or preservation risks for individual file formats. For this purpose, The Library of Congress (LOC)[8] has an extensive collection of resources for a large set of relevant file formats as well as individual assessment of their sustainability factors. For instance, SPSS data files are usually accepted by statistical archives even though it is a binary and proprietary format, since this format has wide adoption and there are (open source) programs and libraries capable of reading and writing this format.[9] While the LOC sustainability factors are highly useful and the assessment of the individual file formats is well elaborated and comprehensive, the information provided is not machine-actionable and thus, we could not build an automated assessment process based on the LOC recommendations.

Cornell University uses a quite similar methodology and provides similar recommendations for their eCommons repository service.[10] In particular, based on a list of criteria, they divide file formats into three categories: 1) High probability for full long-term preservation (e.g. plaint text, PDF/A or PNG), 2) Medium probability for full long-term preservation (e.g. OpenOffice (*.swx), GIF or compressed TIFF) and 3) Low probability for full long-term preservation (WordPerfect (*.wpd) or Microsoft (*.doc)). Even though their list has limited file format coverage and some assessments are disputable, it proved a usable starting point for an initial risk assessment for long-term preservation. From 145 recognized file formats in our data set, 32 were assigned a high probability for successful migration, ten in the medium category and 103 do only have a low probability of successful preservation.

A RESTful Risk Classification Service

The goal of bwDATA Diss project is to preserve dissertation data together with dissertation texts. To be able to assess the preservation risks of datasets, we have implemented and deployed a dataset characterization service, using a simple traffic light visualization, signalling the user the preservation probability of a given file format.

The results of the characterization service can be used either as pre-ingest check, e.g. as a tool for feedback to an initial submission, i.e. flagging unsustainable, unknown or otherwise difficult file formats. Based on this feedback, individual researchers can be advised to re-consider their file format choices (if possible) and their awareness can be raised on the un-sustainability of their format choices. Furthermore, the characterization results may be used to guide a software collection, required to render certain datasets or to prepare an emulation or virtualization strategy.

A characterization request[11] is issued by POSTing a JSON object containing a URL/URI to the dataset and a URL/URI to a preservation policy file. For efficiency reasons we require users to prepare datasets before submission by wrapping the files within an ISO9660 UDF container (CD-ROM / DVD format). This way, a full download of the dataset for characterization is not necessary. Instead, the remote file is mounted and data required for file format characterization is only transferred on request.

8 Sustainability of Digital Formats – Planning for Library of Congress Collections: https://www.loc.gov/preservation/digital/formats/index.html

9 SPSS System Data File Format Family (.sav): https://www.loc.gov/preservation/digital/formats/fdd/fdd000469.shtml

10 Recommended File Formats for eCommons: https://guides.library.cornell.edu/ecommons/formats

11 An example request is explained at: http://classifier.eaas.uni-freiburg.de/

Since user data is only cached in memory, parallel requests can be handled without considering temporary disk space constraints.

Following a characterization request the service will immediately return a session ID, which can be used for querying the status of the characterization request. Depending on the object's size, the characterization may take some time to finish. The requesting client is able to retrieve the characterization result using the session ID. If the characterization is not finished, the client is required to repeat the request later.

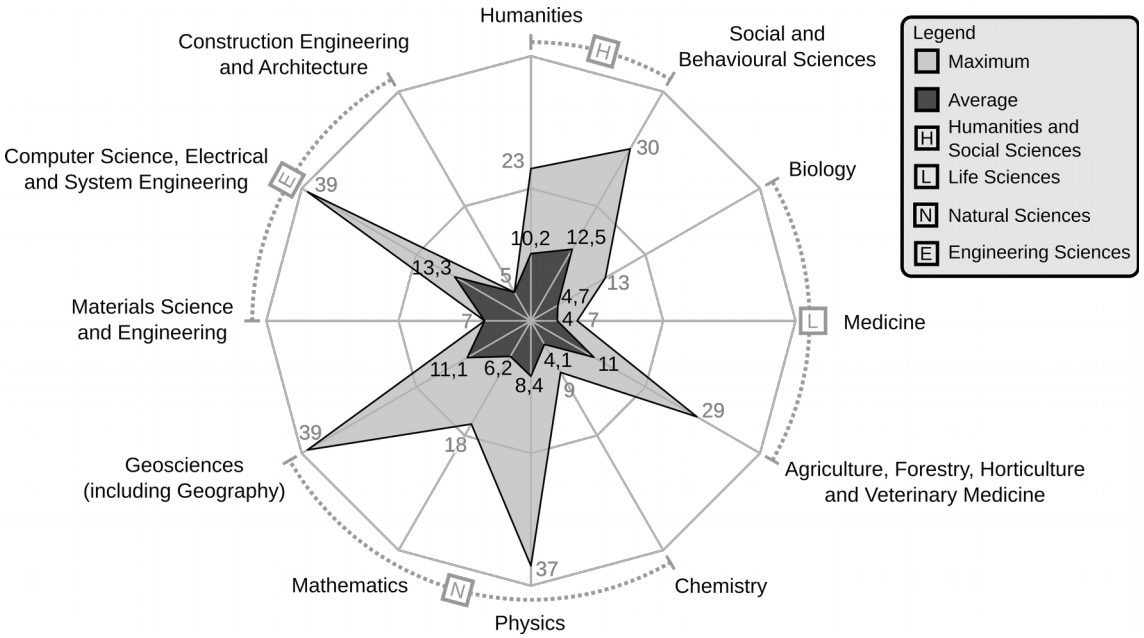

Figure 3. Maximum and average file formats used, grouped by research discipline.

From a Data Centric View Towards a Data Processing View

The classic, migration-driven approach of long-term preservation focuses typically on individual file formats. However, it is likely that research datasets are more heterogeneous, i.e. different file formats are found in a single dataset. To test this hypothesis, we analyzed the average and maximum number of file formats found in a single dataset.

For all datasets, at least four and a maximum of 39 different file formats were found. The variation of average values among the different discipline groups was rather low: on average Humanities and Social Sciences use 11.35 formats, but in particular Life Sciences (6.56), Natural Sciences (7.45) and Engineering Sciences with 8.43 average file formats used are quite similar. Figure 3 visualizes our findings. The numbers determined should not be read as the exact numbers found in each dataset, due to their grouping of file formats and unknown file formats, but should be seen as an approximate lower bound value. However, these results support the argument that some extra attention to the data's software dependencies is necessary. Different files and file formats may have strong interdependencies concerning re-usage and thus preservation planning and preservation actions should take in account these interdependency and aim

for a higher level view as a single file format migration may not be sufficient. Contrary to the heterogeneous nature of the datasets composition, not just 'a' set of software is required for reusing these dataset, but a rather specific or 'homogeneous' software or system setup is required.

Conclusion and Outlook

Our analysis highlighted some technical and conceptual difficulties of keeping research data re-usable. A rather simple file format analysis of research data proved to be a much harder task than anticipated. Tool support was weak and handling a large amount of data is challenging.

If FAIR is the success criteria for successful preservation, a broader and technically more diverse approach is required. A generic research data service can not simply refuse badly rated formats (or datasets containing such files). Such highlighted risk should be the starting point for a productive workflow. For this the data creator should be involved and the potential access and re-use issues of his dataset should be discussed. In some cases the red label is simply due to a failed file format analysis and can be clarified quickly. Furthermore, some (popular/openly documented) file formats can be migrated and tool support exists. However, the analysis also showed that there is a significant portion of problematic datasets where a migration strategy seems not to be an appropriate solution. In this case these datasets pose software dependencies, e.g. software-based runtime environment, which themselves are then subject of their own preservation plans (and problems).

References

Atkinson, M., Gesing, S., Montagnat, J. & Taylor, I. (2017). Scientific workflows: Past, present and future. Elsevier.

Austin, C.C., Brown, S., Humphrey, C., Leahey, A., Webster, P. & Fong, N. (2015). Research data repositories: Review of current features, gap analysis, and recommendations for minimum requirements. *IASSIST Quarterly, 39*(4). doi:10.29173/iq904

Kennan, M.A. & Markauskaite, L. (2015). Research data management practices: A snapshot in time. *International Journal of Digital Curation, 10*(2), 69–95. doi:10.2218/ijdc.v10i2.329

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. I*nformation Services & Use* (Preprint), 1–8. doi:10.3233/ISU-170824

Paul-Stüve, T., Rasch, G. & Lorenz, S. (2015). Ergebnisse der Umfrage zum Umgang mit digitalen Forschungsdaten an der Christian-Albrechts-Universität zu Kiel (2014). Zenodo. doi:10.5281/zenodo.32582

Peng, R.D. (2011). Reproducible research in computational science. *Science, 334*(6060), 1226–1227. doi:10.1126/science.1213847

Press, G. (2016). Cleaning big data: Most time-consuming, least enjoyable data science task, survey says. Forbes, March, 23. Retrieved from https://www.forbes.com/sites/gilpress/2016/03/23/data-preparation-most-time-consuming-least-enjoyable-data-science-task-survey-says

Roche, D.G., Kruuk, L.E., Lanfear, R. & Binning, S.A. (2015). Public data archiving in ecology and evolution: how well are we doing? *PLoS Biology, 13*(11), e1002295. doi:10.1371/journal.pbio.1002295

Simukovic, E., Kindling, M. & Schirmbacher, P. (2013). Umfrage zum Umgang mit digitalen Forschungsdaten an der Humboldt-Universität zu Berlin. doi:10.18452/13568

Tristram, F., Bamberger, P., Cayoglu, U., Hertzer, J., Knopp, J., Kratzke, J., ... Wehrle, D. (2015). Öffentlicher Abschlussbericht von bwFDM-Communities. Retrieved from http://bwfdm.scc.kit.edu/downloads/Abschlussbericht.pdf

Wilkinson, M.D., Dumontier, M., Aalbersberg, I.J., Appleton, G., Axton, M., Baak, A., ... et al. (2016). The FAIR Guiding Principles for scientific data management and stewardship. *Scientific Data, 3*, 160018. doi:10.1038/sdata.2016.18

Woods, K. & Brown, G. (2008). Migration performance for legacy data access. *International Journal of Digital Curation, 3*(2), 74–88. doi:10.2218/ijdc.v3i2.59