Markus, Katharina; Tunnat, Yvonne

Conference Paper — Published Version

ACT NOW, LATE OR NEVER: Make Digital Objects (more) archivable early in their life cycle?

ACT NOW, LATE OR NEVER:

Make Digital Objects (more) archivable early in their life cycle?

Katharina Markus
ZB MED – Information Centre for Life Sciences
Germany
Markus@zbmed.de

Yvonne Tunnat
ZBW – Leibniz Information Centre for Economics
Germany
y.tunnat@zbw.eu

Abstract – Newly acquired or published objects might be corrupt or might not conform to the archive's best practices. In some cases the library or archive even has to ask the data provider for replacements. The advantage of a pre-archival workflow to detect or prevent problems early in institutional data processing is depicted in this paper.

Keywords – Archivability, Digital Preservation, Validity, PDF format

Conference Topics – Exchange; Innovation

I. INTRODUCTION

When archives obtain objects and prepare them for ingest into the archival system, they follow digital preservation best practices. The archive department usually is responsible for the object preparation step which is sometimes conducted at a time significantly after the institution obtained the objects. But to consider best practices and to conduct this step earlier, e. g. directly after acquisition, might save time and curation effort later on. It may also allow preservation of information that is lost otherwise. This paper will introduce two use cases at the institutions *ZBW – Leibniz Information Centre for Economics* and *ZB MED – Information Centre for Life Sciences*. The institutions obtain control over objects relevant for this paper at two processing stages – publication (ZB MED) and acquisition (ZBW). This paper analyses in a qualitative manner the benefit of introducing preservation best practices into the early processing steps of publication and acquisition. The analysis is based on an implemented new workflow (ZBW) and implementation planning (ZB MED). It can serve as a basis for other institutions in similar situations where the archives deal with high amounts of objects that also require relatively high amounts of curation.

II. BACKGROUND AND RELATED WORK

The preservation community defined general best practices for preservation [1]–[4]. One example for best practices used as quality criteria for objects is shown as follows: The German National Library (DNB) defines five different ingest levels, which increase the quality of files regarding preservation with each level from data integrity through identification of file formats, unrestricted access to files for DNB, available technical metadata and, finally, to valid files according to format validation. The DNB conducts quality checks during ingest and rejects files if integrity is not provided and formats are not identifiable [5].

Curating objects according to preservation best practices during transfer to archives may result in huge curation efforts for archives and archive departments, stalling objects in the pre-ingest or ingest step [6]. Efforts might be due to obtaining necessary rights [7], [8], dealing with non-standard and inconsistent infrastructure and data structure as well as missing files [6], [8], [9] or simply defective data [10]. Personal communication of the author Yvonne Tunnat with various members of the digital preservation community shortly after publishing a blog post regarding curation efforts and relevant tools used during acquisition shows interest in this topic as well.

Increasing conformance with these best practices was termed for this paper as increased *archivability* [11], which was defined by Banos and Manolopoulos

18th International Conference on Digital Preservation
iPres 2022, Glasgow, Scotland.

[11] as "whether [a website] has the potential to be archived with completeness and accuracy" but is used in this context for objects in a broader sense. The processing of objects to make them more archivable, like detecting and repairing defective files, were defined as *actions increasing archivability*. Institutions can conduct these actions during any step of the object processing workflow (from object creation until archiving, see fig. 1). For this paper, the authors divided processing steps and actions into those which are part of a *pre-archive* workflow (WF) and those that are part of the archive WF, where specifically pre-ingest and ingest steps are located. In the pre-archive WF, departments other than the archive are responsible for the object processing and actions are localized earlier in the processing (see fig. 1).

As far as the authors were able to determine, literature rarely analyses which archivability increasing actions are best conducted in early processing steps by departments other than the archive. Preservation best practices are implicitly targeted at archive departments and the archives are recommended to take "an active role in [digital information's] maintenance early in its life cycle" [12]. Still, the authors found the prospect of early curation actions mentioned in the context of digitization projects [13]–[15], research data [12] and web archiving [11]. Skinner and Schultz [13] address digitized objects but also consider born-digital material. They devote a chapter to preservation best practices for digitized objects as part of creation and acquisition, in which they recommend the set-up of an inventory, the definition and documentation of recommended file formats, metadata and data structures, the generation of checksums and establishment of explicit permission to preserve the objects.

Selected best practices for ZB MED and ZBW with relevance for this paper are similar: consistent data structure, recommended file formats, among them PDF/A with embedded open fonts, valid objects and metadata standards established in the research community. The actions that increase archivability are related to these best practices. The authors assume that introducing these actions in early object processing steps results in a reduction of total curation effort for the institution (see fig. 1).

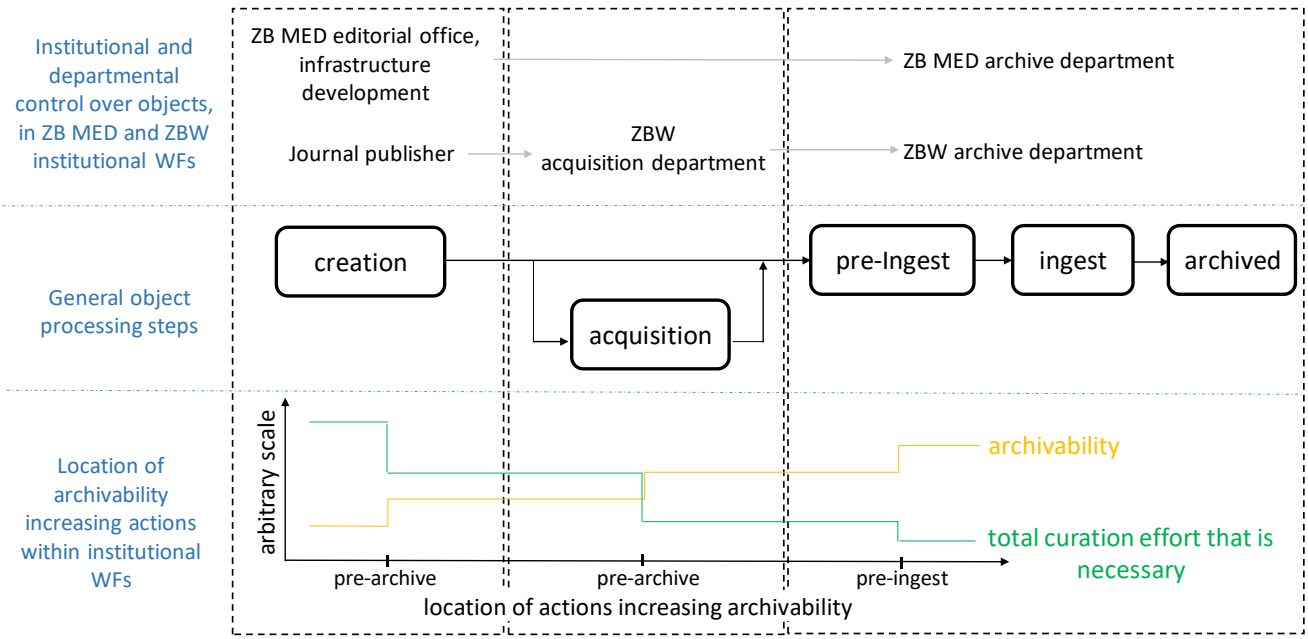

Figure 1 Depicted are object processing steps (from object creation to archived), the respective steps where ZB MED and ZBW obtain control of the objects, as well as three possible locations of archivability increasing actions (pre-archive during publishing, pre-archive during acquisition or pre-ingest within the archive department). The diagram shows the result of conducted archivability increasing actions as an increase in archivability and a decrease in total curation effort. Increase and decrease have been determined as a result from qualitative analysis instead of quantitative measurements and therefore the scale is arbitrary.

18th International Conference on Digital Preservation
iPres 2022, Glasgow, Scotland.

III. Method

To answer the research question "Is generating archivable digital objects early in their life cycle worth the (staff and process development) effort?", this short paper uses the method "actual practice" (as opposed to best practice, as examinations about this sub-topic in Digital Preservation seem to be rarely addressed in literature), analyzing available literature (chapter II) and two uses cases: description and evaluation of ZB MED's pre-archival WF (chapter IV. A) and ZBW's pre-archival object processing (chapter IV. B). In the referenced blog posts the used tools are described and more practical points of the workflow are examined. This paper uses a qualitative analysis regarding benefits of early archivability actions and assesses the impact on the objects which have to be archived.

IV. Use Cases

A. ZB MED Use Case

ZB MED provides several publication services to the life science community, among them the PUBLISSO Gold publication portal [16] and, in collaboration with the Association of the Scientific Medical Societies in Germany, the German Medical Science (GMS) publication portal [17]. Publications on these platforms are intended to be archived in ZB MED's own archive, which is a separate system. Since a high number of data sets are archived retroactively, at a significantly later time than their publication date, various challenges became apparent when the archive collected data sets from the GMS portal. Still, ZB MED has control over formal quality assurance (QA) of publications on the platforms. Accordingly, it can incorporate various suggestions from its own archive department into publication processes and infrastructure in order to increase archivability. A summary of possible actions improving archivability during publication steps follows. These are in various states of implementation.

A well-known challenge for archiving is clearing necessary rights [7], [8]. In addition to rights of objects, fonts can also be copyrighted which hinders embedding during PDF/A migration. Using open fonts during publication allows later embedding in PDF/A without reviewing and verifying the usage rights of the used font. In case the font used during publication does not allow embedding by the archive, the institution can resort to another font that does. But this change in fonts may lead to changes in content display, which requires additional QA, and therefore effort from the archive. Optimally, PDF/A with already embedded fonts is used for publication.

A significant challenge during ingest relates to the publication's data structure. If the structure is inconsistent, the institution preparing the data packages for the archive (data provider or the archive itself) cannot rely on an entirely automated workflow. Instead, it needs to identify exceptions and map the different data structures of provided objects to the archive's data structure. In the established workflow at ZB MED, about 0.1% publications contained exceptions that resulted in additional handling. While changes to data structures over time are probably unavoidable, the publisher may help with later data transfer by documenting a data structure schema as well as exceptions and new versions alongside respective objects. This may be useful not only for a transfer into archives but also for exit scenarios.

At the object level, recommendations of file formats that are more or less suited for digital preservation are well known in the digital preservation community [18]. Close collaboration with editorial offices helps with communicating these as best practices to authors. Additionally, the introduction of validation and a documentation of publication versions might also offer opportunities: The benefit of pre-archive validation is detailed in the ZBW Use Case (see chapter IV. B). Versioning of identifiers in metadata when new versions of a publication are generated allows for automated or partly automated update workflows connecting platform and archive. This, in turn, should reduce efforts of communicating updates between staff of different departments while also decreasing risks of human error.

Going beyond the purely technical level, markup languages can also serve as metadata standards as part of the object itself. As text publications are not necessarily restricted to the PDF format but become increasingly reusable for machines when published as XML, selecting subject-specific markup languages according to preservation best practices becomes relevant as well. Examples of well-known subject-specific markup languages are bioschemas [19] or MathML [20]. Recommending these standards for publications with preservation best practices in mind while also consulting the scientific community can evolve into a new task for subject-specific archives. In case of machine readability of molecular

18th International Conference on Digital Preservation
iPres 2022, Glasgow, Scotland.

structures of chemical compounds, ZB MED researched open, well-used and maintained markup languages. They consulted FAIRsharing [21], taking into account referenced maintainers, number of databases that use the standard and whether the standard is open. They investigated usage of the standard in popular software used in the research community, for which they referred to people with a background in chemistry and related fields. A preliminary selection resulted in openSMILES [22] as the preferred standard.

In general, integrating the above-mentioned recommendations into object creation processes is expected to reduce total curation efforts as an automated object transfer is enhanced and likelihood for later handling is decreased. For further work, ZB MED attempts the implementation of the mentioned suggestions as far as technically possible.

B. ZBW Use Case

The ZBW – Leibniz Information Centre for Economics provides digital documents like articles and research papers on its many presentation platforms like EconStor [23] or other instances which are all available via EconBiz [24].

The ZBW established Digital Preservation in 2015 to ensure long-term availability for their hosted content. The Digital Preservation Archive is a dark archive, based on the System Rosetta developed by Ex Libris [25]. All the content is presented to the users by other representation platforms, mostly based on DSpace [26].

However, Digital Preservation is the last step in the object processing pipeline, just as it is at ZB MED (see chapter IV. A). For most workflows that presents no problem, as the material on the DSpace platform is published immediately after acquisition and the ingest is done the night after.

For objects acquired under National and Alliance Licences, though, the hosting on ZBW servers and therefore the ingest to the Rosetta archive is done months or even years after the acquisition of the material.

After such a long time, the data providers (usually publishers like Emerald, De Gruyter and Elsevier) have long since moved on to other projects, so that it is time-consuming and sometimes impossible to get a replacement if parts of the data are missing or corrupted.

Therefore, the ZBW staff responsible for the acquisition has established an automated preliminary data check workflow pre-archive. The publishers deliver the data, in most cases a large amount of PDF files, in Zip folders.

During the past years, the pre-archive workflow included:

- Unpacking the zipped files
- Integrity check (via checksums)
- Completeness check

Newly implemented into the pre-archive workflow are:

- Checking for password protection (which would impede data migration)
- Running the PDF files through tools to check for errors

The tools used are Grep, PDFinfo and, mainly: ExifTool [27]. The workflow in detail, the implementation of the workflow, the staff time used for daily work and the handling of different ExifTools error messages are described in detail in an OPF blog post published in February 2022 [28].

Tests have shown that certain error messages hint at the PDFs not being archivable, sometimes not even accessible for the users. For those, the ZBW acquisition department can ask for a replacement directly after acquisition. As many PDF files are password-protected, the ZBW rights department and the data providers have agreed to delete the password-protection. To accomplish this, the ZBW acquisition department has set up another automated workflow.

As only open source tools are used, the invested resources are calculated as curation effort, specifically as staff time of the involved departments acquisition and archive.This includes:

- copying the PDF files to the hotfolder where the tools conduct their actions
- preliminary judgment of the findings (especially if a new error occurs, which has not been evaluated yet)
- if a new error occurs, the ZBW archive department checks if the affected PDF files can be migrated to PDF/A-2b
- if a new error occurs, the acquisition department performs a manual check to see if the PDF is accessible. This is also done for some errors, such as “PDF header not at beginning of the file”.

The curation effort for a bulk of 1,000 PDFs for the newly established workflow, in average, requires an

18th International Conference on Digital Preservation
iPres 2022, Glasgow, Scotland.

hour of staff time. This includes error-handling and asking for replacement when a PDF file is corrupted. This workflow now takes up more time during acquisition due to additional actions that aim for better archivability.
The ingest into the archive, in comparison, is now fully automated and usually does not need extra staff time. Only if errors occur does the ZBW archiving department have to work on these and spend staff time. Nevertheless, the new WF is worth the extra time during acquisition, as corrupted data is detected early and can be replaced, whereas it would be permanently lost to the archive or in general otherwise. No matter how good the digital curation workflows are: if the data is too corrupted to be repaired or even lacking contents to begin with, there is nothing to be done about it at a later stage. Either the contact to the publisher has gone cold, so that the ZBW acquisition team cannot get hold of the data provider and thus, the object anymore. Or the ZBW and the publisher negotiated that the contents can be hosted (and thus archived) when the data is no longer available from the publisher's websites. In this case, if defective data is discovered a significant time after archiving and the publisher does not provide it anymore either, the content is lost for good.
As a side effect: The data providers have so far been grateful for the information about corrupt files, as they also want to offer a high data quality on their platforms for their users.
The ZBW staff established these workflows quite recently. In the future the acquisition department will evaluate the workflows regularly and, if necessary, extend or alter them.

C. Tools

While ingesting the data into the ZBW Rosetta Archive, several tools are used: DROID, JHOVE, NLNZ Metadata Extractor, just to name the most important. These tools extract technical metadata like the file format including the format version, detect password-protected files and identify basic information about size, creation date and a lot of other information useful to ensure long term-availability.

As a side effect, the archive department usually detects files that are not accessible or otherwise corrupted.

During the pre-archive workflow after acquisition, the acquisition department uses Grep, PDFinfo and ExifTool (see chapter IV. B). The usage of the tools is regularly evaluated, e. g. via tool benchmarking; comparing which tool is best suited for a certain task, mostly with regard to file validation. This has been done thoroughly for the file formats:

- TIFF [29]
- JPEG [30]
- GIF [31]
- PDF [32].

As tools and their usage frequently evolve, close preservation watch is essential. For instance, in December 2017, when the ZBW archive department examined the validation tools for PDF, ExifTool was not even considered, although it would have been of use back then. ZBW staff did not include it in the evaluation only because they did not yet know about the tool.

For some use cases, tools could also be inappropriate, as they take too long, give too many false alarms (false positives) or their validation is too thorough for pre-archive needs, like JHOVE for PDF [33].

The tools ZB MED uses for preparation of objects for archiving in its present workflow are a self-developed Submission Application (SubApp) as well as JHOVE and veraPDF in pre-ingest processing. The archive department is responsible for operating these. During the subsequent ingest the archive department uses further tools, the same as ZBW (see above) which are not detailed here. The SubApp generates data packages and detects exceptions in the data structure, whereas JHOVE so far detected invalid image files during pre-ingest processing. Exceptions and invalid files require individual processing by the archive and the editorial office, as part of the otherwise automated pre-ingest workflow. The archive department is in close contact with the publishing platforms regarding analysis and evaluation of tools and changes to objects and WFs.

V. Findings and Summary

As shown in the use cases, ZBW and ZBW identified several actions which, when implemented in pre-archive workflows, may reduce curation efforts presently or in the future. As ZB MED detects various exceptions during pre-ingest with their present WF, they expect better automation if data structure is documented early in a stringent way. As additional opportunities, ZB MED identified the use of open

18th International Conference on Digital Preservation
iPres 2022, Glasgow, Scotland.

fonts for PDF publications, markup languages suited for preservation as well as documenting object versions in a standardized way. ZBW discovered corrupted data well after acquisition with their old WF. The new WF contains validation with ExifTool pre-archive, as part of acquisition. This allows early detection of invalid, password-protected and corrupted files and subsequent exchange of files when contact to the provider is still established. With these analyses, the authors expand on the recommendation by Skinner and Schultz [13] with specific tools (ExifTool) and proposed implementations (e. g. open fonts for PDF publications) based on actual practice.

Both institutions come to the conclusion that early incorporation of these best practices, tools and actions seems to prevent significantly higher efforts later. "Later" meaning here, if archivability increasing actions are conducted a significant time after publication or acquisition. The reasons are twofold: first, when the institution is still in contact with an external data provider, obtaining correct versions of files (corrupt, invalid) and clarifying rights (password protected) requires less effort than re-establishing contact months or years after data provision was concluded. Additionally, at this point in time the department sometimes can still obtain data that would be lost to the archive otherwise. Secondly, the departments involved in publishing already process objects at an individual level. Additional curation at that stage requires less effort than stopping automated archiving processes later on. Nonetheless, not every curation action can or should be implemented in pre-archival WFs. Therefore, the archive departments selected the above-mentioned actions and best practices in exchange with editorial offices and the acquisition department. They maintain contact with these pre-archive departments as well, re-evaluating workflows while also taking organizational and technical conditions into account. Still, this evaluation of prospective and actual implementations described here might help with the scaling of archiving workflows, not just for the institutions mentioned in this paper but for others as well, because all are faced with increasing amounts of all kinds of materials that need to be archived.

References

[1] J. Mitcham and P. Wheatley, "Digital Preservation Coalition Rapid Assessment Model (DPC RAM) (Version 2.0)." [Online]. Available: http://doi.org/10.7207/dpcram21-02 (accessed Feb. 14, 2022)

[2] A. Beking *et al.*, "Digital Curation Decision Guide," *NDSA Publ.*, Dec. 2020, doi: 10.17605/OSF.IO/Q8C47. (accessed Feb. 14, 2022)

[3] E. Faulder *et al.*, "Digital Processing Framework," report, Aug. 2018. Accessed: Jan. 18, 2022. [Online]. Available: https://ecommons.cornell.edu/handle/1813/57659

[4] CoreTrustSeal Standards and Certification Board, "CoreTrustSeal Trustworthy Data Repositories Requirements 2020–2022," Nov. 2019, doi: 10.5281/zenodo.3638211. (accessed Feb. 14, 2022)

[5] Deutsche Nationalbibliothek, "Spezifikation der Dateiformat-Policy für die Sammlung von Netzpublikationen der Deutschen Nationalbibliothek. Version 1.0." Oct. 24, 2012. [Online]. Available: http://nbn-resolving.de/urn:nbn:de:101-2012102408 (accessed Feb. 14, 2022)

[6] S. Morrissey and A. Kirchhoff, "Managing Preservation Costs with managed Ingest: The Portico Straight-to-Ingest Project," in *Proceedings of 16th International Conference on Digital Preservation*, Amsterdam, Sep. 2019, pp. 317–322. doi: 10.17605/OSF.IO/VW7RJ. (accessed Feb. 14, 2022)

[7] Rosenthal, David, "Why Preserve E-Journals? To Preserve The Record," Jun. 10, 2007. https://blog.dshr.org/2007/06/why-preserve-e-journals-to-preserve.html (accessed Dec. 26, 2021).

[8] B. Sprout and M. Jordan, "Distributed digital preservation: preserving open journal systems content in the PKP PN," *Digit. Libr. Perspect.*, vol. 34, no. 4, pp. 246–261, Jan. 2018, doi: 10.1108/DLP-11-2017-0043. (accessed Feb. 14, 2022)

[9] R. Arora, M. Esteva, and J. Trelogan, "Leveraging High Performance Computing for Managing Large and Evolving Data Collections," *Int. J. Digit. Curation*, vol. 9, no. 2, Art. pp. 17-27, Oct. 2014, doi: 10.2218/ijdc.v9i2.331. (accessed Feb. 14, 2022)

[10] B. Sprout and S. Romkey, "A Persistent Digital Collections Strategy for UBC Library," in *Proceedings of the Memory of the World in the Digital Age: Digitization and Preservation Conference*, Vancouver, British Columbia, Canada, 2013, pp. 256–268. [Online]. Available: http://www.unesco.org/new/fileadmin/MULTIMEDIA/HQ/CI/CI/pdf/mow/VC_Sprout_Romkey_26_B_1540.pdf (accessed Feb. 14, 2022)

[11] V. Banos and Y. Manolopoulos, "A quantitative approach to evaluate Website Archivability using the CLEAR+ method," *Int. J. Digit. Libr.*, vol. 17, no. 2, pp. 119–141, Jun. 2016, doi: 10.1007/s00799-015-0144-4. (accessed Feb. 14, 2022)

[12] D. R. Bleakly, "Long-term spatial data preservation and archiving: What are the issues," *Sand Rep. SAND 2002*, vol. 107, 2002, doi: 10.2172/793225. (accessed Feb. 14, 2022)

[13] K. Skinner and M. Schultz, *Guidelines for Digital Newspaper Preservation Readiness*. Atlanta, GA: University of North Texas Libraries, UNT Digital Library, https://digital.library.unt.edu, 2014. Accessed: Jan. 07, 2022. [Online]. Available: https://digital.library.unt.edu/ark:/67531/metadc282586/

[14] K. Dohe and R. Pike, "Integration of Project Management Techniques in Digital Projects," in *Project Management in the Library Workplace*, vol. 38, Emerald Publishing Limited, 2018, pp. 151–166. doi: 10.1108/S0732-067120180000038013. (accessed Feb. 14, 2022)

[15] Deutsche Forschungsgesellschaft, "12.151 DFG-Praxisregeln 'Digitalisierung' [12/16]." Accessed: Feb. 10, 2022. [Online].

Available: http://www.dfg.de/formulare/12_151/12_151_de.pdf (accessed Feb. 14, 2022)

[16] "PUBLISSO Gold Open Access publication portal." https://www.publisso.de/en/publishing/ (accessed Feb. 21, 2022).

[17] "German Medical Science publication portal." https://www.egms.de/dynamic/en/index.htm (accessed Feb. 21, 2022).

[18] C. Arms and C. Fleischhauer, "Digital Formats: Factors for Sustainability, Functionality, and Quality," *Arch. Conf.*, vol. 2005, no. 1, pp. 222–227, Jan. 2005.

[19] A. J. Gray, C. A. Goble, and R. Jimenez, "Bioschemas: from potato salad to protein annotation," presented at the International Semantic Web Conference (Posters, Demos & Industry Tracks), 2017.

[20] R. Miner, "The importance of MathML to mathematics communication," *Not. AMS*, vol. 52, no. 5, pp. 532–538, 2005.

[21] S.-A. Sansone *et al.*, "FAIRsharing as a community approach to standards, repositories and policies," *Nat. Biotechnol.*, vol. 37, no. 4, Art. pp. 358-367, Apr. 2019, doi: 10.1038/s41587-019-0080-8. (accessed Feb. 14, 2022)

[22] "Simplified Molecular Input Line Entry Specification Format (SMILES)." https://doi.org/10.25504/FAIRsharing.qv4b3c (accessed Feb. 14, 2022)

[23] ZBW, "EconStor." https://www.econstor.eu/ (accessed Feb. 14, 2022).

[24] ZBW, "EconBiz." https://www.econbiz.de/ (accessed Feb. 14, 2022).

[25] Ex Libris, "Rosetta." https://exlibrisgroup.com/products/rosetta-digital-asset-management-and-preservation/ (accessed Feb. 14, 2022).

[26] dSPACE, "DSpace." https://www.dspace.com/en/pub/home.cfm (accessed Feb. 14, 2022).

[27] Harvey, Phil, "ExifTool." https://exiftool.org/ (accessed Feb. 14, 2022).

[28] Tunnat, Yvonne, "Validation with ExifTool: Quick and not so dirty," *Open Preservation Foundation Blogs*, Feb. 04, 2011. https://openpreservation.org/blogs/pdf-validation-with-exiftool-quick-and-not-so-dirty/?q=1 (accessed Feb. 14, 2022).

[29] Tunnat, Yvonne, "TIFF format validation: easy-peasy?," *Open Preservation Foundation Blogs*, Jan. 17, 2017. https://openpreservation.org/blogs/tiff-format-validation-easy-peasy/ (accessed Feb. 14, 2022).

[30] Tunnat, Yvonne, "Error detection of JPEG files with JHOVE and Bad Peggy – so who's the real Sherlock Holmes here?," *Open Preservation Foundation Blogs*, Nov. 29, 2016. https://openpreservation.org/blogs/jpegvalidation/ (accessed Feb. 14, 2022).

[31] Tunnat, Yvonne, "Good GIF hunting: JHOVE's GIF validation skills," *Open Preservation Blogs*, Dec. 05, 2017. https://openpreservation.org/blogs/good-gif-hunting/ (accessed Feb. 14, 2022).

[32] Tunnat, Yvonne, "JHOVE – the one and only PDF validator," *Open Preservation Foundation Blogs*, Dec. 19, 2017. https://openpreservation.org/blogs/jhove-the-one-and-only-pdf-validator/ (accessed Feb. 14, 2022).

[33] M. Lindlar, C. Wilson, and Tunnat, Yvonne, "A PDF Test-Set for Well-Formedness Validation in JHOVE - The Good, the Bad and the Ugly," in *iPRES Japan*, Japan, Kyoto, Sep. 2017, p. 11. Accessed: Feb. 21, 2022. [Online]. Available: https://files.dnb.de/nestor/weitere/ipres2017.pdf