10th Conference on Grey Literature and Repositories: proceedings [online]. Prague: National Library of Technology, 2017. ISSN 2336-5021. Available from: **http://nrgl.techlib.cz/conference/conference-proceedings/**.

FROM THE DISSEMINATION OF ELECTRONIC THESES AND DISSERTATIONS TO THEIR LONG-TERM ARCHIVING

Eliška Pavlásková

eliska.pavlaskova@ruk.cuni.cz

Institute of history and Archive of Charles University, Czech Republic

Abstract

Since 2006 it has been mandatory for Czech universities to make electronic theses and dissertations accessible on the Internet. Nevertheless, theses and dissertations are also historical archival materials of fundamental historical value, and need to be treated as such. In the year 2016, the Archive of Charles University initiated a change of the current policy on thesis submission. The emphasis was on using formats specifically suitable for long-term preservation (the format PDF/A, in particular). The objective was to collect theses in a form which may discontinue the practice of submitting printed versions and facilitate the use of electronic versions as the original archival materials. The presentation focuses on the historical development of thesis collection (including an analysis of files submitted during the years 2006-2016), submission policy description, and its implementation into the submission process.

Keywords

Electronic Theses and Dissertations; Digital Preservation; Archiving; PDF/A; Format Policy

Introduction

In general, theses or dissertations are materials with a significant value for the history of science and culture. As an outcome of university education and (mainly in the case of doctoral dissertations) research, these resources have a lasting value and significance and therefore constitute a heritage that should be protected and preserved for current and future generations.

Today, theses and dissertations are created and processed mostly in electronic form. Digital institutional repositories drastically change the ways in which theses are accessed, disseminated, and internally processed. Electronic versions of theses and dissertations (ETDs) include texts, databases, still and moving images, audio, graphics, software, and web pages, among a wide and growing range of formats.

Digital preservation has turned into a pressing challenge for institutions with the obligation to preserve digital objects over years. It is the collective term for actions that will ensure access to digital content in the future. The method of preservation is defined by a philosophical and practical understanding of the digital content (Digital Preservation Strategy, 2011). The Institute of the History of Charles University and Archive of Charles University is responsible for the long-term preservation of theses and dissertations in analogue (paper) form. With the advent of ETDs, their curation has become an obligation of the Archive as well. The first step in planning and executing the digital preservation strategy is the formulation and implementation of a new format policy. The policy was formed with regard for the needs of students and digital preservation and with practices internationally recognized as being the best, and it takes the recommendations of the National Archives into consideration.

Background

Since 2006, Charles University has been accepting electronic versions of student theses and storing them in an institutional repository. By 2010, all students were required to deposit their ETDs via the web interface of the Student Information System (SIS). SIS creates a simple submission information package for ingest into the institutional repository, and it provides a mechanism for the identification and validation of deposited files. This policy is sufficient for dissemination and access to ETDs. Nevertheless, theses and dissertations are also historical archival materials. Until now, theses and dissertation have been archived in physical form (mostly on paper). Students of Charles University finish approximately 17 000 theses and dissertations every year and storage of physical materials become uneconomical and impractical. Archiving of digital data instead of paper is logical but not simple solution.

At Charles University, theses and dissertations are considered as archival materials under Act No. 499/2004 Coll., on archive and record management. There are several possible ways in which digital archival materials can be handled in compliance with the Act. Nevertheless, all of the variants demand that the archives have the ability to create submission information packages (SIPs) according to the structural and formal rules set by the National Archives. Consequently, any format policy issued by Charles University needs to take the recommendations of the National Archives into consideration.

Development of the format policy

The format policy of Charles University is influenced by the concrete demands of Czech archival legislation. As a specialized archive under Act No. 499/2004 Coll., on archive and record management, the Archive of Charles University must follow the recommendations of the National Archives and their National Digital Archive (NDA) project. The NDA distinguishes between three groups of formats (Bernas, 2009):

- Preferred formats (e. g. plain text, XML, CSV, TIFF, Wave…)
- Accepted formats (e.g. PDF, JPEG2000, GIF…)
- Formats with a low durability (e.g. MS Word, internal formats of graphical applications…)

For still text and image documents, government resolution no. 1338 of 3 November 2008 demands the use of PDF/A-1a (ISO 19005-1 – Portable Document Format – Electronic document file format for long-term preservation), PNG (ISO/IEC 15948:2004 - Portable Network Graphics) and TIFF (Tagged Image File Format - revision 6 - Uncompressed) for use as output formats of electronic record management systems (ERMS) (Bernas, 2009).

The electronic theses and dissertations at Charles University are not managed by EMRS, so the resolution requirements are not mandatory. Nevertheless, the university should take account of them while creating its format policy.

The international community of digital preservation experts can base format assessment and policy creation on factors and categories used by leading institutions in the field. However, the British Library advises caution in the use of pre-existing documents. “Published guidelines, policies and assessments have a ripple effect and are often reused without considering the underlying evidence or the influence of unique organizational requirements. Meta assessments that make recommendations based on surveys of what other organizations do add a further level of obfuscation.” (Pennock et al., 2014)

Table 1 - Format evaluation factorsTable 1 summarizes the criteria used in format evaluation and assessment by the British Library (Pennock et al, 2014), the Library of Congress (Sustainability Factors, 2017), MIT Libraries (File Formats for Long-term Access), and the National Library of the Netherlands (Rog a Van Wijk, 2009). The factors described influence the feasibility and cost of preserving information content in the face of future change in the technological environment in which users and archiving institutions operate.

British Library	**Library of Congress**	**MIT libraries**	**National Library of the Netherlands**
Documentation and Guidance	Disclosure	Open, documented standard	Openness
Adoption and Usage	Adoption	Common usage by research community	Adoption
Complexity	Transparency	Standard representation (ASCII, Unicode) Unencrypted Uncompressed	Complexity
	Self-documentation		Self-documentation
External Dependencies	External dependencies		Dependencies
Legal Issues	Impact of patents	Non-proprietary	
Technical Protection Mechanisms	Technical protection mechanisms		Technical Protection Mechanism (DRM)
Development Status			
Software Support			
Embedded or Attached Content			
Other Preservation Risks			
			Robustness

Table 1 - Format evaluation factors

The characteristics mentioned above should be used as a theoretical basis for the process of choosing preferred formats for theses and especially their annexes. Nevertheless, the selection of file formats for ETDs should be considered in the wider strategic context of Charles University, its Archive, its digital preservation needs, and its abilities. As mentioned in the Digital Preservation handbook: “At all times, the answer to digital preservation issues is not to try and “do everything”. Your strategy ought to move you towards simple and practical actions rather than trying to support more file formats than you need. (File formats and standards, 2017)”. From a practical point of view, it is crucial to reduce the complexity of data collection and storage only in necessary formats.

Preliminary analysis

As mentioned above, Charles University has been collecting theses and dissertations for several years. ETDs were collected in PDF version 1.3 or newer, and students were allowed to deposit also an annex in the form of a single file or ZIP archive. Only version of the PDF and the occurrence of text information in the file were validated. Students were provided with no

additional guidelines regarding the file creation. Deposited files represent a large set of files which allow us to run format analyses on a significant number of relevant objects.

Preliminary analyses of deposited files had two main objectives – to gain knowledge about the formats of annexes and to test identification tools. The analysis of PDFs with the main text of the theses was not relevant, as students were not required to deposit specific version of the PDF and we assume a change of policy from PDF 1.3 to PDF/A.

We tested a set of 481,396 files submitted as annexes of 2,528 theses and deposited between January 2015 and February 2016. The analysis identified 148 different formats consisting of 174 puids (PRONOM unique identifiers[1]). The distribution of formats in the set takes the form of standard "long tail" distribution. A large group of formats is represented only by the single file.

Table 2 shows the distribution of formats with an occurrence higher than 1 %. The most common format is the plain text format, which represents the source code of computer programs in most cases.

Format	Occurrence
Plain Text File	38.70 %
Portable Network Graphics	12.28 %
[not able to identify]	9.00 %
Hypertext Markup Language	6.78 %
JavaScript file	6.23 %
JPEG File Interchange Format	4.98 %
Extensible Markup Language	3.90 %
Java language source code file	2.52 %
Extensible Hypertext Markup Language	2.04 %
Apple Double Resource Fork	1.70 %
ZIP Format	1.26 %
Acrobat PDF 1.5 - Portable Document Format	1.18 %
Windows Portable Executable	1.04 %

Table 2 - The most frequent formats

[1] Unique identifiersof the file format developed by the National Archives (GB).

Approximately 80 % of ETDs with annexes (more than 2,000) has only one file attached as an annex. Figure 2 - Files per ETD (annexes including 10 or more files)shows the distribution of files per ETD for annexes including ten or more files. It is again “long tail” distribution. The largest sets of files are related to a small number of ETDs.

As a conclusion of the preliminary analysis, we determined that the format control in annexes must be flexible. There was a need for format guidelines in the case of standard format groups (e.g. images or video) and for a policy enabling students to deposit complex software objects with unknown characteristics.
Fido² was chosen as the main identification tool - the reasons for this decision include the demands placed on server administration and implementation requirements on the part of SIS.

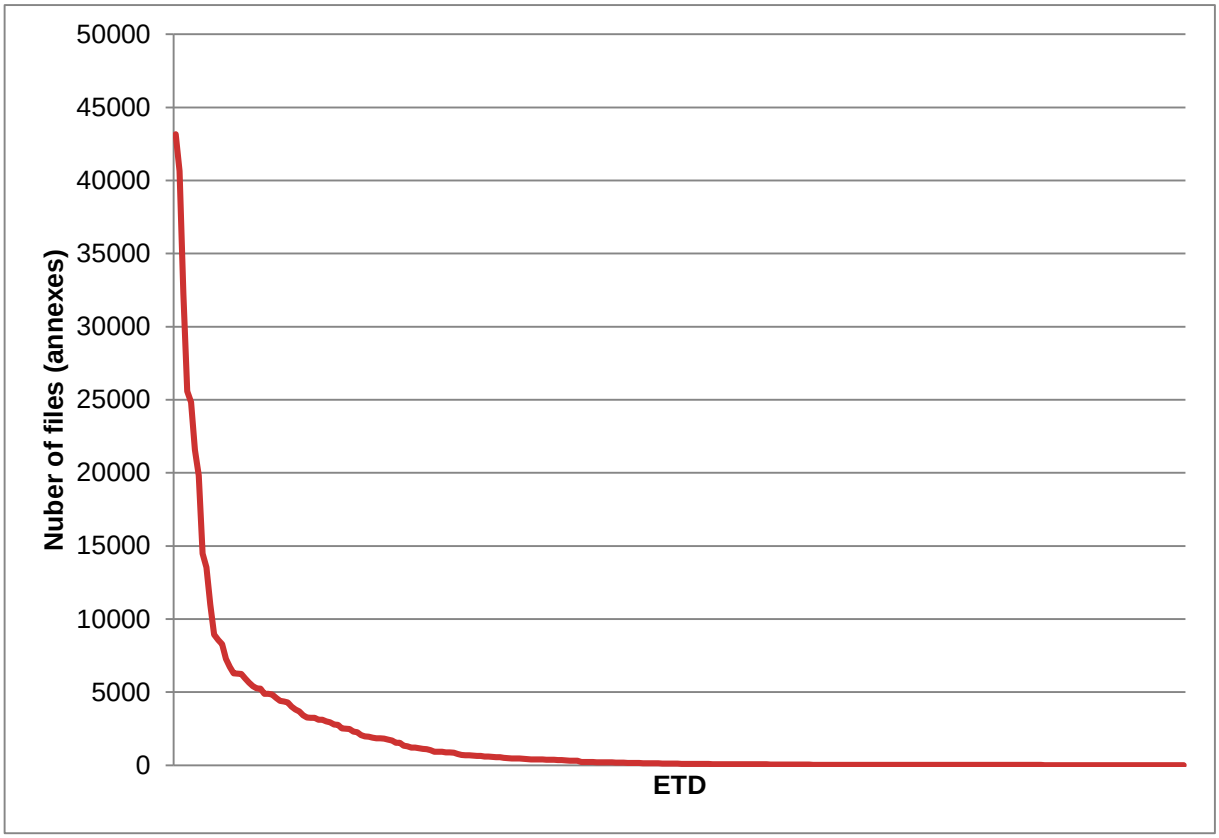

Figure 1 - Files per ETD (annexes including 10 or more files)

Structure of ETD

A thesis is usually perceived as one text document. However, from the point of view of the university administration, an ETD is a complex object consisting of several parts with different characteristics and needs in terms of the format used. The detail structure of the ETD is described by figure 2.

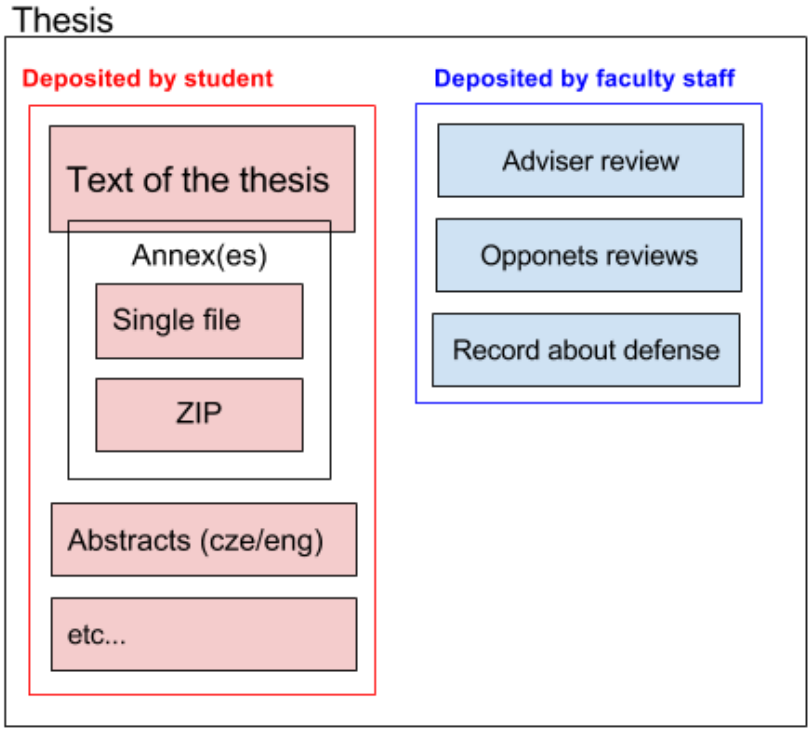

Figure 2 - Structure of the ETD

The text of the thesis is the main part of the complex object. It is created by a student, and it has textual character. It is almost always a born-digital object, and common word editors or document preparation systems are used for its creation. There is no feasible way to accomplish a trustworthy conversion from the format used by the student to create the thesis to the format preferred for archiving. The content of the text is heterogeneous, and it is essential to check the output of the conversion manually, otherwise the content of the document can be significantly changed or even destroyed. For example, mathematical equations are frequently difficult to convert without errors and need to be checked by an expert in the field (ideally by their author).

An obvious choice for a text document is the PDF/A format. It is recommended by the National Archives as well as by the majority of leading institution in the field (see, for example, Rimkus, 2014). The word processing software that is common in the academic community (Microsoft Word, Libre and Open Office) is usually able to create a PDF/A format[2]. The Archive of Charles University contacted study departments of faculties with a request for further information on the software used. The most important response came from the Faculty of Mathematics and Physics with regard to documents created by the typesetting system TeX. The faculty provided the archive with a template for creatiing PDF/A 2u from TeX.

Given the problematic nature of PDF/A version 3 (Wheatley, 2015), only versions 1 and 2 were allowed. As regards the National Archives recommendation, level “a” was chosen and level “u” was later added on the request of the Faculty of Mathematics and Physics. Level "a" and level "u" both enforce the use of correct unicode mapping. Moreover, level "a" demands the use document structure.

[2] With exception of Pages and Word for Mac.

Annex(es) are a file or files created by students as an appendix to the main text. In textual form, it may be part of the main text. The inclusion of text annexes into the main text is encouraged, but in some cases, it is useful to use a separate file. The content of the annex can be a set of images, sounds, a video, tables, or computer programs. In some cases, annexes consist of research data and should be treated with consideration of their reuse. The size of the annex object can be significant.

Based on recommendations of the National Archives and after consultations with the Czech National Film Archives, the following formats were chosen for individual content types:

Text annex(es):

- PDF/A (version 1a or 2u)

Image annex(es) - the use of PDF/A (version 1a or 2u) is strongly recommended to students to for image annexes. If not possible, JPEG format can be used.
Audio annex(es):

- Waveform audio format (WAV, *.wav nebo *.wave)
- Moving Picture Experts Group Phase Audio Layer III (.mp3)

Audio-visual (video) annex(es):

- Moving Picture Experts Group Phase 2 (MPEG-2, *.vob)
- Moving Picture Experts Group Phase 4 (MPEG-4, *.mp4)

Annex(es) with character of data (tables):

- Comma-separated values (CSV, *.csv)
- Extensible Markup language (XML, *.xml) – the submission package must contain the relevant XSD or DTD
- Plain text file (*.txt)

The formats listed above have the status of an allowed format. All other formats are considered as non-approved. The exception was created with regard for the scientific or research data, software, computer application or simulations. The submission must be accompanied by an application containing a simple description of the data characteristics.

Other documents – errata, abstracts, summaries and thesis proposition (in Figure 2 as "etc.") are mostly textual[3] objects used for administrative purposes. From a format point of view, they have the same (even simpler) character as the text of the thesis. Therefore, PDF/A (version 1a or 2u) was chosen.

Reviews and record about defence – set of materials in textual form, created usually by the faculty staff. It can be born-digital, or in some cases, digitalized. The output formats of scanners and equipment used by staff need to be taken into consideration when making a format policy. PDF/A (version 1a or 2u) was chosen for the pilot stage of implementation.

[3] Charles University is considering the use of non-textual video files as abstracts in sign language.

Pilot stage of implementation, analysis and policy changes

The pilot stage for the implementation of the new format policy started in March 2017 and ended in June 2017. During this period, we were able to collect a set of more than 3,000 ETDs and more than 5,000 files submitted as annexes.

Students were provided with an information site[4] containing provisional guidelines for the creation of PDF/A and basic guidelines for submission of annexes. An electronic help desk was created for answering specific problems with the ETD submission.

During the pilot period, we identified several areas that need to be improved or customized. We encountered problems regarding the behaviour of the validation tool, problems with PDF/A conversion in word processing software, and errors in the workflow for annex processing.

As mention above, we use the open source software VeraPDF[5] as a validation tool. VeraPDF is currently the only existing open source PDF validation tool, and it is able to validate all versions of PDF/A against a set of rules based on the PDF/A specification (ISO 19005). We also created our own version of the validation profile (a set of rules used for validation).

The second challenging area was user behaviour and the use of an information site with guidelines. During the whole pilot period, we were constantly analysing user queries in the help desk application and updating the information site with guidelines. From a format policy point of view, the most serious problems were caused by conversion in word processing or typesetting software. Approximately 11 % of queries concerned Unicode mapping in Microsoft Word (all versions). Nevertheless, a student used glyphs with no representation in Unicode only in one extremely specific case. In all other cases, an error occurred during file conversion. The conversion errors were usually independent of the software (Microsoft Word) and font versions used. The most common problem was the use of "□" as a bullet point. Approximately 3 % of problems reported by students were caused by processing transparency in images or graphs. We developed strategies to avoid this type of error and published them as a part of the guidelines. Table 3 shows the total number of users' queries.

[4] **https://www.cuni.cz/UK-7987.html**
[5] **http://verapdf.org/**

Contens of the query	**Occurence**
Misunderstanding of guidelines	66
Unclear queries (student did not react to request for more information)	33
Errors of interface (timeouts, etc.)	33
LaTeX (transferred to MFF)	32
Non-relevant to format policy (e.g. requests to change the title)	31
Unicode mapping	28
Vera 1.4 - error in profile (critical failure in system, eliminated after two hours)	18
Use of Office 2007	14
Submission form for annexes and its use	14
Use of Pages or Office for Mac	11
Processing of transparency	10
Digitalized reviews	9
Validation profile malfunction	5
Obsolete guidelines (from the website of the faculty)	3
Misuse of Adobe Acrobat	2
Indesign	1
Request for additional information	1
Personal opinion about PDF/A	1
Total number of queries	312

Table 3 - Users' queries analysis

Specific set of problems constitute a typesetting system TeX. We closely collaborate with the Faculty of Mathematics and Physics, where we were able to find an expert with knowledge and experience in the use of TeX.

The number of files attached as annexes to 75 ETDs totalled 5,834 files, and 53 different file formats were identified. Ninety percent of the files were attached to only two ETDs. Unfortunately, this distribution prevents us from carrying out a reliable analysis of the formats used. It is safe to assume that the authors of both ETDs use formats specific to their work. There is, therefore, no way to interpret the data correctly.

During the pilot stage, we also encountered numerous problems with documents deposited by the faculty staff (reviews and records about defence). It was decided that the forced use of

PDF/A for these types of documents will be stopped and that the practice of archiving them in analogue form as part of the student's file will be preserved.

Conclusion

The long-term preservation of ETDs at Charles University can be done only on the condition of an existing and preserved format policy. Different approaches must be chosen for the main text of the thesis, annexes and supplementary documents. According to the practices internationally recognized as the best and according to Czech legal requirements, PDF/A is probably the only possible choice for submitted texts. Viable implementation of PDF/A collection must be based on format validation and comprehensive guidelines for students.

The format policy regarding annexes should be flexible and enable the submission of large and heterogeneous sets of files. Two possible ways of submission were facilitated - submission of files in allowed formats and submission of non-approved files supplemented with short additional information about the data deposited.

Supplementary materials such as reviews or record of defence should be part of the student file in an analogue form. Alternatively, digitalization equipment with the ability to produce PDF/A should be provided for the administrative staff of Charles University.

References

BERNAS, Jiří. Národní digitální archiv. *Knihovna* [online]. 2009, **20**(1), p. 22-29 [Accessed 16 September 2017]. ISSN 1802-8772. Available from: **http://knihovna.nkp.cz/knihovna91/bernas.htm**.

Digital Preservation Strategy [online]. Wellington: Archives New Zealand Te Rua Mahara o te Kāwanatanga: National Library of New Zealand Te Puna Mātauranga o Aotearoa. 2011 [Accessed 25 September 2017]. Available from: **http://archives.govt.nz/sites/default/files/Digital_Preservation_Strategy.pdf**

File formats and standards*. Digital Preservation Handbook* [online]. Glasgow: Digital Preservation Coalition, 2017 [Accessed 23 September 2017]. Available from: **http://www.dpconline.org/handbook/technical-solutions-and-tools/file-formats-and-standards**

File Formats for Long-term Access. MIT Libraries [online]. Cambridge (MA): Massachusetts [Accessed 2 October 2017]. Available from: **https://libraries.mit.edu/data-management/store/formats/**

MCGUINNESS, Rebecca, Carl WILSON, Duff JOHNSON and Boris DOUBROV. VeraPDF: open source PDF/A validation through pragmatic partnership. In: *14th International Conference on Digital Preservation* [online]. [Accessed 23 September 2017]. Available from: **https://ipres2017.jp/wp-content/uploads/28Rebecca-McGuinness.pdf**

PENNOCK, Maureen, WHEATLEY, P., MAY, P. Sustainability assessments at the British Library: Formats, frameworks and findings. In: *Proceedings of the 11th International Conference on Digital Preservation*. 2014. p. 141-148. Available also from: **https://fedora.phaidra.univie.ac.at/fedora/get/o:378110/bdef:Content/get**

RIMKUS, Kyle, Thomas PADILLA, Tracy POPP and Greer MARTIN. Digital Preservation File Format Policies of ARL Member Libraries: An Analysis. *D-Lib Magazine* [online]. 2014, **20**(3/4). [Accessed 23 September 2017]. DOI: 10.1045/march2014-rimkus. ISSN 1082-9873. Available from: **http://www.dlib.org/dlib/march14/rimkus/03rimkus.html**

ROG, Judith; VAN WIJK, Caroline. Evaluating file formats for long-term preservation. *Data Analysis and Knowledge Discovery*, 2008, 24.1: p. 83-90. Available also from: **https://www.kb.nl/sites/default/files/docs/KB_file_format_evaluation_method_27022008.pdf**

Sustainability Factors. *Sustainability of Digital Formats: Planning for Library of Congress Collections* [online]. Washington: Library of Congress, 2017 [Accessed 23 September 2017]. Available from: **https://www.loc.gov/preservation/digital/formats/sustain/sustain.shtml**

WHEATLEY, Paul, Peter MAY, Maureen PENNOCK and Simon WHIBLEY. *PDF Format Preservation Assesment* [online]. Version 1.3. London: British Library, 2015 [Accessed 23 September 2017]. Available from: **http://wiki.dpconline.org/images/e/e8/PDF_Assessment_v1.3.pdf**