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

Mgr. Eliška Pavlásková, Ph.D.
Institute of history and Archive of Charles University

10th Conference on Grey Literature and Repositories, 19. 10. 2017

O elektroaffinitě aluminia.

Jaroslav Heyrovský.

Úvod a rozvržení.

Původním cílem této práce byl pokus o zjednodušení výroby čistého hydroxydu aluminia zavedením podobným Solvayově při výrobě louhu (návrh prof. F. J. Domana). Že by se totiž aluminium z roztoku na rtuťové kathodě vylučovalo, a takto vzniklá aluminiová amalgáma nechala se oxydovat, dávajíc přímo čistý hydroxyd.

Za tím účelem bylo v první řadě nutno seznati, jak dalece je amalgámová elektroda aluminiová vůbec schopna existence a je-li to elektroda zvratná.

Studie elektrolytického potenciálu této elektrody však předpokládá znalost koncentračních poměrů mezi ionty roztoků solí aluminiových. Proto jsem vyšetřil měřením elektromotorických sil

Background

- ETDs at Charles University
  - *2006 – ETDs accepted in the digital repository*
  - *2010 – submission via web interface of the Student Information System (SIS)*
    - Submission workflow – SIP creation

- Act No. 499/2004 Coll. on Archiving and Records Management
  - *Preservation of digital archival materials*
  - *Recommondation of the National Archives*

Digital Preservation Strategy
Format Policy - Sources

- Specific needs of Charles University
  - *Designated community*
  - *Producers – students and staff*
  - ***Preliminary analysis** of submitted ETDs*
- National Digital Archive recomondations
  - *Preferred formats*
  - *Accepted formats*
  - *Formats with low durability (e.g. MS Word, internal formats of graphical applications…)*
- Best practices

Best Practices
Format Evaluation Factors

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

Preliminary Analysis (annexes)

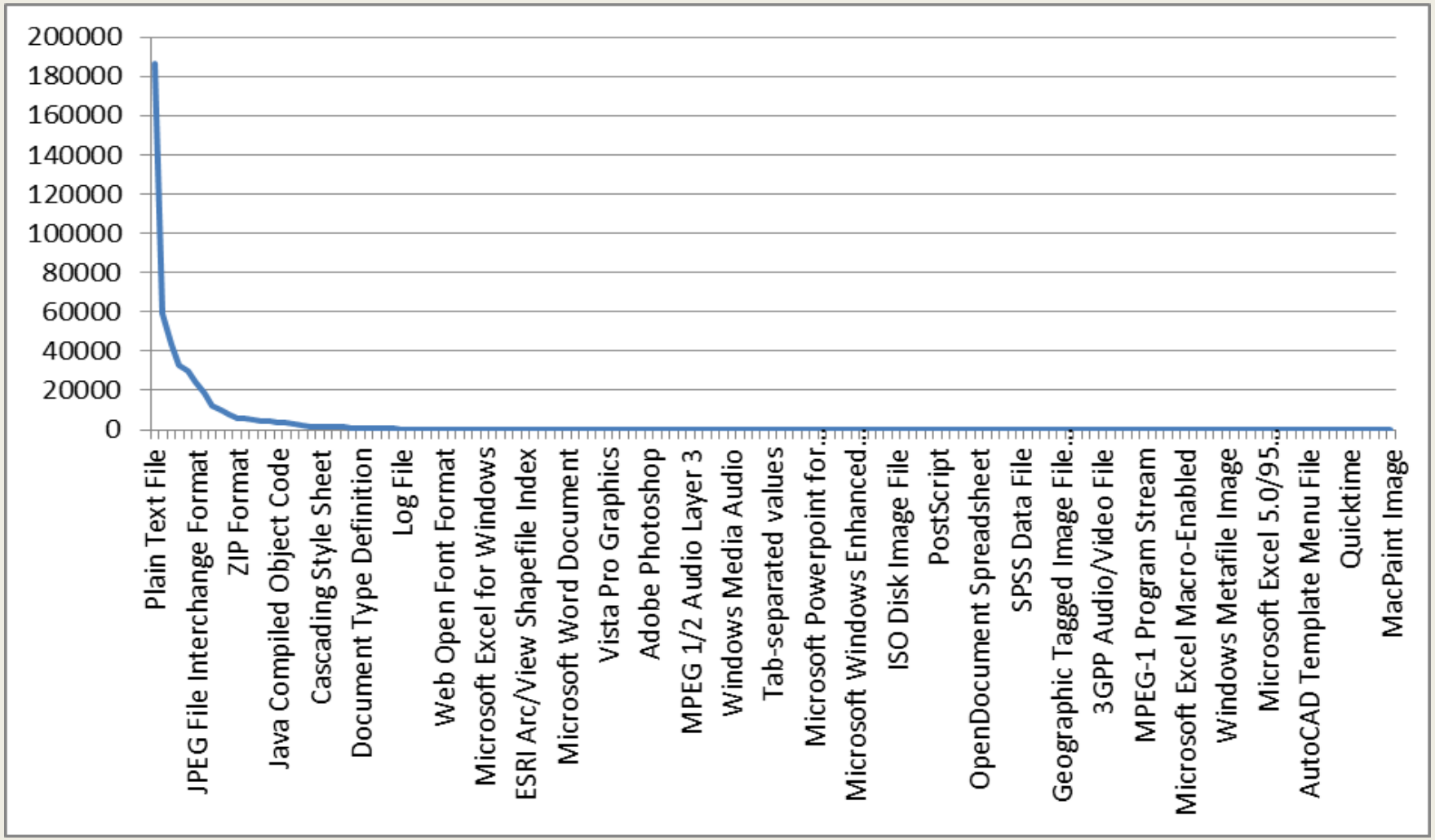
- January 2015 - February 2016

- 2 528 thesis
- 481 396 files

- 148 different formats (174 puids)
- More than 20 image or graphic formats
- More than 10 audiovisual formats

Preliminary Analysis

Distribution of Files per Format



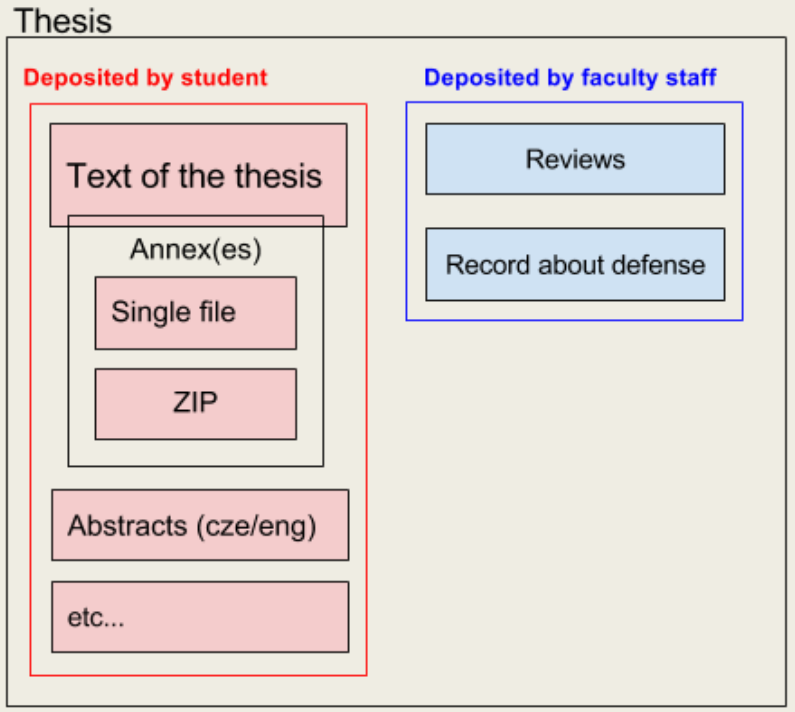
Preliminary Analysis

The Most Frequent Formats

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%

Format Policy

Structure of ETD



Text of the Thesis

- Born-digital
  - *Typesetting system TeX*
  - *Word processors*
- Manual check of conversion is necessary

- PDF/A – 1a, 2u
  - *PDF/A 3 – preservation risk*

Format Policy

Annex(es) - Allowed Formats

- Text annex(es):
  - *PDF/A (verze 1a or 2u)*
- Image annex(es)
  - *Joint Photographic Experts Group File Interchange Format (JPEG/JFIF, přípony: .jpeg, .jpg)*
  - *PDF/A (verze 1a or 2u)*
- Audio annex(es):
  - *Waveform audio format (WAV, *.wav or *.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 languge (XML, *.xml) – submission package must contain relevant XSD or DTD*
  - *Plain text file (*.txt)*

Format Policy

Annex(es) - Non-approved Formats

Accopanied with application:

a) Title of the Thesis

b) Author ‘s name

c) Reasons for the application

d) Format of the attachment (including version or other specifications) or program language (including version)

e) Applications necessary for rendering or use of attachment(s)

f) Operation system (including version)

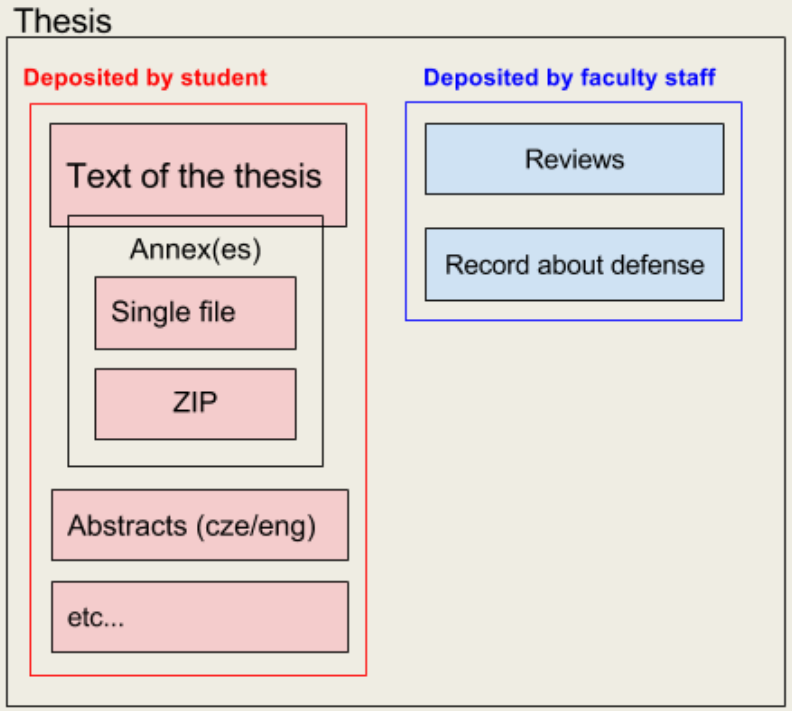
g) License information

h) List of submitted files

i) Any other relevant information (for example list and description of used modules and libraries)

Format Policy

Structure of ETD



Abstracts etc.

- PDF/A – 1a, 2u

Reviews and record about defence

- Born-digital or digitized
- PDF/A – 1a, 2u

Pilot Stage of Implementation

- Submission workflow modifications
  - *VeraPDF – validation tool*
  - *Fido – identification tool*
  - *Application form for description of annexes in non-approved formats.*
- Information site
  - *Provisional guidelines for PDF/A conversion*
  - *Basic guidelines for annex submission*
- Electronic help desk

- March 2017 – June 2017
- Cca 3 000 ETDs and 5 000 files submitted as annexes.

Challenging areas
Submission interface

- VeraPDF
  - *Active development– communication with development team*
  - *Inconsistent behavior*
    - Versions
    - Parsers (PDFBox x Greenfield)
  - *Processing errors*
- Custom validation profile – set of validation rules
- Submission workflow
  - *Processing errors*
  - *Processing of Annexes*

Challenging areas
PDF/A Creation

- LaTeX
- Unicode mapping
  - *Mostly error during conversion*
- Processing of transparency
  - *Images*
  - *Graph*
- Understanding of the guidelines

- PDF/A version
  - *MS Word 2007 – PDF/A 1b*
  - *MS Word 2010, 2013 – PDF/A 1a*
  - *MS Word 2016 – PDF/A 3a*
  - *Libre Office – PDF/A 1a*
  - *Adobe Acrobat Professional*

  - *PDFCreator?*

Pilot Stage Assement

- Submission workflow optimalization
  - *Asynchronous validation*
  - *Validation output report*
- Portable Network Graphics (png) allowed for image annexes
- Changes in the application form for submission of an annex in non-allowed format
- Documents deposited by faculty staff in PDF

THANK YOU FOR YOUR ATTENTION

eliska.pavlaskova@ruk.cuni.cz