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

This presentation is licensed under the Creative Commons: <u>CC-BY-SA-4.0</u>, via http://www.nusl.cz/ntk/nusl-367304



O elektroaffinité aluminia.

Jaroslav Heyrovský.

č, 923

Uvod a rozvržení.

Původním cilem této práce byl potus o sjednodušení výroby čistélio lydrovyde aluminia savedením podobným Solvayově při výrobě louhu (návch prof. F. J. Domana). Že by se totie aluminium s rostoku na rtutové kathodě vylučovalo, a takto vzniklá aluminiová amalgáma nechala se ozydovat, dávajíc přímo čistý hydrozyd.

La tim riclem bylo v promi rade mutno sernat, jak dalece je amalgamora elektroda aluminiova vibec schopma existence a je-li to elektroda soratud.

Studie elektrolytického poleuciálu této elektrody však předpokládá znalost koncentračních poměrů mezi iouly rostoků soli aluminiových. Proto jseu 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

| 200000 | | | | | | | | | | | | | | | | | | | | | | | | |
|--------|---|------------------------------|------------------|-----------------------|--------------------------|----------------------|-----------------------|-------------------------------------|--------------------|-----------------|------------------------|---------------------|---|-----------------------------|---------------------|--------------------------|----------------|-------------------------------|---|-------------------------------|--|----------------------------|-----------|----------------|
| | | | | | | | | | | | | | | | | | | | | | | | | _ |
| 180000 | | | | | | | | | | | | | | | | | | | | | | | | _ |
| 160000 | + | | | | | | | | | | | | | | | | | | | | | | | _ |
| 140000 | + | | | | | | | | | | | | | | | | | | | | | | | _ |
| 120000 | + | | | | | | | | | | | | | | | | | | | | | | | _ |
| 100000 | + | | | | | | | | | | | | | | | | | | | | | | | _ |
| 80000 | + | | | | | | | | | | | | | | | | | | | | | | | _ |
| 60000 | + | | | | | | | | | | | | | | | | | | | | | | | _ |
| 40000 | + | | | | | | | | | | | | | | | | | | | | | | | _ |
| 20000 | + | $\mathbf{\lambda}$ | | | | | | | | | | | | | | | | | | | | | | _ |
| 0 | +++++++++++++++++++++++++++++++++++++++ | | | | | | | 11111 | 11111 | 11111 | | | | 1111 | | 111111 | | | | 1 | | | 11111 | |
| | <u>e</u> | at at | de la | et | uo e | Format | ٧S | č č | ics | do | č | <u>.</u> ≘ | es. | ed. | ile i | et E | <u>e</u> | ie ie | ∎Ę | ed | 92 Ge | e. e | ne | 89 |
| | Ť | Ige Format | 8 | she | Ë j | | þ | pd 8 | l d | ŝ. | ye | . vu | nt f | nc | ч С | readsheet | ъ | ы с | r s | lde | 2 Ja | 2 5 | Quicktime | дa |
| | ĕ | Ъц | 2 ರ | e S | i <u></u> ≣ _ | 3 2 | Ē. | | i a | g | Ľa | ч. | <u> </u> | ha | e e | ğ | at | ě d | st g | Ľ. | - <u>-</u> | e u | <u>.</u> | = |
| | Ē | e e | <u> </u> | ~ | ě | ÷ | ~ | Ξċ | 5 0 | 0 | - | | | | | | | | ≚ ~ | | | | - | Ē |
| | | - 00 - | | 4 | | 드 | - | 50 | ίā | Ę | . <u>e</u> | 5. | n te | Ъ | <u></u> <u></u> | Ζã | 5 | <u> </u> | ≂ ≽ | 6 | j≣ e | Σ | ď | |
| | a. | ang | - 9 - 9 | St | ЗеГ | Fon | for | ape | , or | e Ph | udio | Medi | werp | vsEr | m % | Spre: | PSS D | in line | io/vi gram | scro- | etafil Exce | te M | α. | 2 |
| | Plain Text File | chang | ed Ob | ing Stv | Гуре Г | en Fon | el for \ | Shape! | ta Pro (| obe Ph | 2 Audio | /s Medi | :parate(Powerp | lows Er | Disk Im | nt Sprea | SPSS Data File | gged Im | udio/VI rogram | Macro- | Metafilo oft Exce | olate M | ď . | lacha |
| | Plai | terchang | piled Ob | ading Sty | nt Type D | Dpen Fon | xcel for \ | w Shapet | /ista Pro | Adobe Ph | 1/2 Audio | ows Medi | -separated oft Powerp | indows Er | io Disk Im | ru nent Sprea | SPSS D | Tagged Im | L Program | el Macro- | vs Metafilo osoft Exce | mplate M | а : : | MacPaint Image |
| | Plai | Interchang | ompiled Ob | ascading Sty | nent Type D | b Open Fon | t Excel for Windows | /iew Shapet | Vista Pro Graphics | Adobe Photoshop | G 1/2 Audio | ndows Medi | ab-separated soft Powerp | Windows Er | ISO Disk Image File | ru sument Sprea | SPSS D | ic Tagged Im | ьР Audio/VI G-1 Program | xcel Macro- | lows Metafile icrosoft Exce | Template M | ď : | МасРа |
| | Plai | ile Interchang | Compiled O | Cascading Style Sheet | ument Type D | Veb Open Fon | soft Excel for \ | c/View Shapet | Vista Pro | Adobe Ph | IPEG 1/2 Audio | Windows Media Audio | Tab-separated values crosoft Powerpoint for | oft Windows Er | ISO Disk Im | ru Jocument Sprea | SPSS D | phic Tagged Im | sure Audio/Video File PEG-1 Program Stream | ft Excel Macro- | indows Metafile Image Microsoft Excel 5.0/95. | VD Template M | ď : | MacPa |
| | Plai | 5 File Interchang | ua Compiled Ob | Cascading Sty | ocument Type D | Web Open Font Format | rosoft Excel for \ | Arc/View Shapet licrocoft Word D | Vista Pro | Adobe Ph | MPEG 1/2 Audio Layer 3 | Windows Medi | Tab-separated values Microsoft Powerpoint for. | osoft Windows Er | ISO Disk Im | ru enDocument Sprea | SPSS D | graphic Tagged Im | אטרא אטמוס/ אטפס דופ MPEG-1 Program Stream | soft Excel Macro- | Windows Metafile Image Microsoft Excel 5.0/95 | CAD Template M | a : | MacPa |
| | Plai | PEG File Interchang | Java Compiled Ob | Cascading Sty | Document Type Definition | Web Open Fon | Aicrosoft Excel for \ | RI Arc/View Shapefile Index | Vista Pro | Adobe Ph | MPEG 1/2 Audio | Windows Medi | Tab-separated Microsoft Powerp | icrosoft Windows Er | ISO Disk Im | ru penDocument Sprea | SPSS D | eographic Tagged Im | 3GPP Audio/V MPEG-1 Program | crosoft Excel Macro- | Windows Metafile Microsoft Exce | utoCAD Template M | a : | MacPa |
| | Plai | JPEG File Interchange Format | Java Compiled Ob | Cascading Sty | Document Type E | Web Open Fon | Microsoft Excel for V | ESRI Arc/View Shapefile Index | Vista Pro | Adobe Ph | MPEG 1/2 Audio | Windows Medi | Tab-separated Microsoft Powerp | Microsoft Windows Enhanced. | ISO Disk Im | OpenDocument Spreadsheet | SPSS | Geographic Tagged Image File. | אט Audio/ Vi MPEG-1 Program | Microsoft Excel Macro-Enabled | Windows Metafil Microsoft Exce | AutoCAD Template Menu File | g . : | MacPa |

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

| Thesis | |
|----------------------|----------------------------|
| Deposited by student | Deposited by faculty staff |
| Text of the thesis | Reviews |
| Annex(es) | Record about defense |
| Single file | |
| ZIP | |
| Abstracts (cze/eng) | |
| etc | |

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

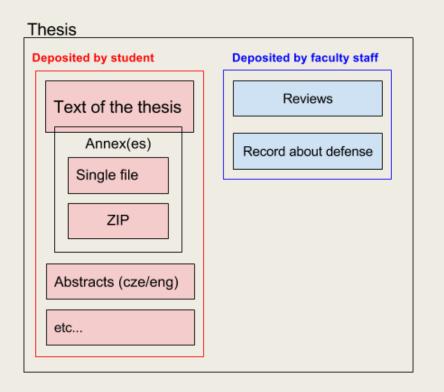
- Text annex(es):
 - PDF/A (verze 1a or 2u)
- Image annex(es)
 - Joint Photographic Experts Group File Interchange Format (JPEG/JFIF, přípony: .jpeg, .jpg)
 - PDF/A (verze 1a or 2u)
- Audio annex(es):
 - Waveform audio format (WAV, *.wav or *.wave)
 - Moving Picture Experts Group Phase Audio Layer III (.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