Prototyping the archiving of spreadsheets

Asbjørn Skødt, Danish National Archives

Open Preservation Foundation, Demonar, December 2022

What did I prototype

Programmatic solutions to:

- Converting spreadsheets to the file formats XLSX Strict and ODS
- Comparing the contents of original to conversion
- Changing and removing content from spreadsheets to meet our file format policies
- Validating spreadsheets
  - Test existing validators for the standards
  - Develop new validator for the format policies

Why did I prototype

- Test and confirm the feasibility of existing libraries and tools for our proposed new file format policies
- Very specifically finding a solution to handling conversion, changing and validation of the **Strict** variant of XLSX file format
- Support advancing from an analytical phase to an implementation phase for ingesting XLSX Strict and/or ODS in our collections

You can find all the prototypes on Github

- **CLISC**
  Primary demonstration prototype, CLI tool
- **Archsheerary**
  NuGet library to build your own spreadsheets application
- **convert-spreadsheet**
  Workflow / CLI conversion tool
- **validate-spreadsheet**
  Workflow / CLI validation tool
- **GUISC**
  GUI to CLISC, work in progress
- **ODS-ArchivalRequirements**
  ODS and OOXML implementation in Java, work in progress

Languages, libraries and tools I used

Handling XLSX Strict

- C#
- Visual Studio 2022
- **Open XML SDK**
- **Excel Interop** and Excel
- CommandLineParser
- LibreOffice
- Beyond Compare 4

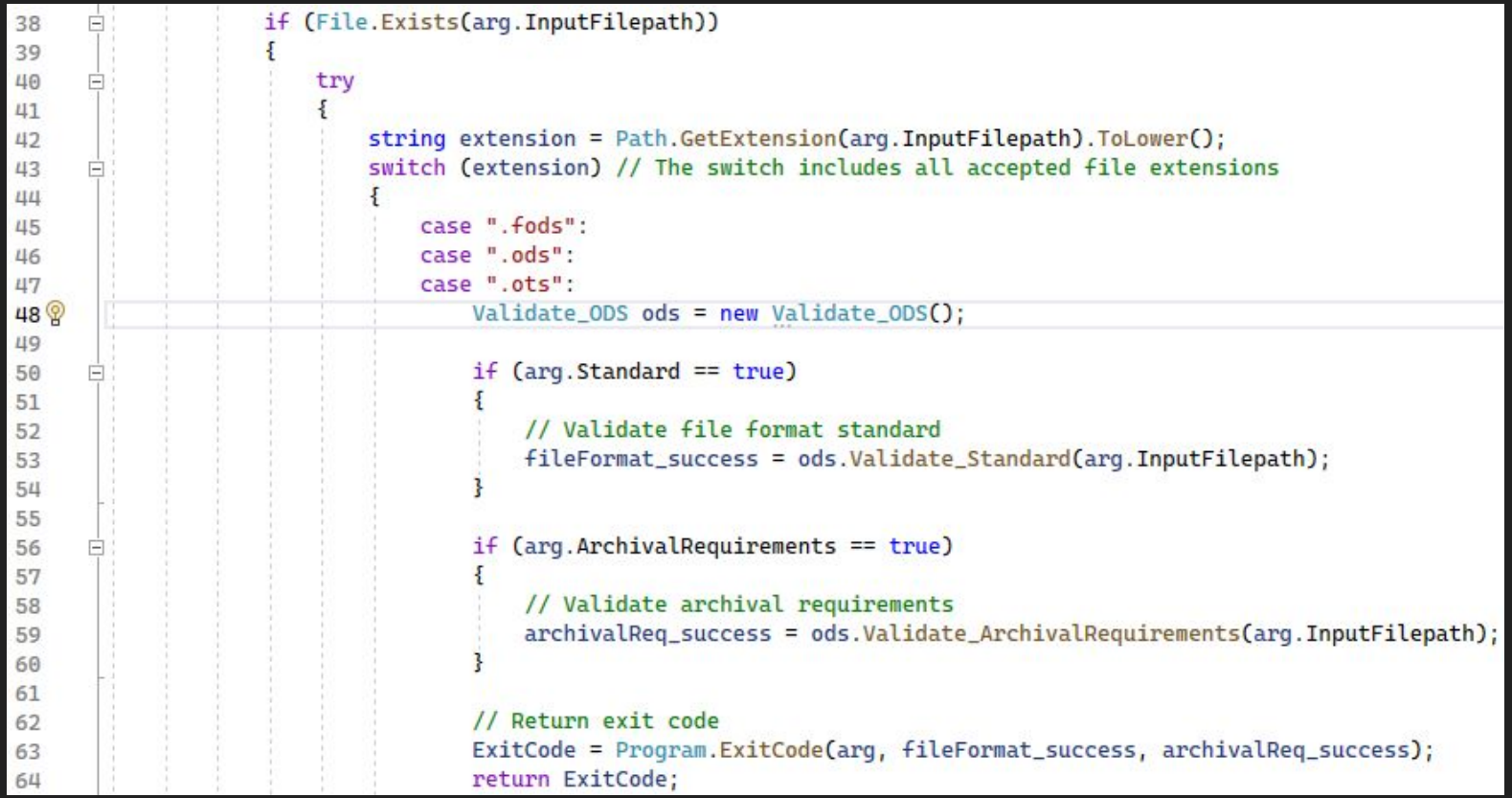
Handling ODS (and XLSX)

- Java
- IntelliJ IDEA
- **ODF Toolkit**
- ODF Validator
- LibreOffice
- Apache Commons
- **Apache POI**
- OOXML Validator

Most of the software is open-source

Other tools

- **stackoverflow**, GitHub, Google



```
38          if (File.Exists(arg.InputFilepath))
39          {
40              try
41              {
42                  string extension = Path.GetExtension(arg.InputFilepath).ToLower();
43                  switch (extension) // The switch includes all accepted file extensions
44                  {
45                      case ".fods":
46                      case ".ods":
47                      case ".ots":
48                          Validate_ODS ods = new Validate_ODS();
49
50                          if (arg.Standard == true)
51                          {
52                              // Validate file format standard
53                              fileFormat_success = ods.Validate_Standard(arg.InputFilepath);
54                          }
55
56                          if (arg.ArchivalRequirements == true)
57                          {
58                              // Validate archival requirements
59                              archivalReq_success = ods.Validate_ArchivalRequirements(arg.InputFilepath);
60                          }
61
62                          // Return exit code
63                          ExitCode = Program.ExitCode(arg, fileFormat_success, archivalReq_success);
64                          return ExitCode;
```

Code example of OpenDcoument Spreadsheets validation from *validate-spreadsheet* tool

Results

- No programmatic approach, other than Excel Interop (even this has errors), exists when converting to XLSX Strict - This **needs** to be developed
- Converting between XLSX and ODS using LibreOffice results in nearperfect conversions based on visual inspections
- I did not find or program a proper tool for content comparison
- Changing and removing content in/from XLSX Strict is feasible - All our file format policy requirements were met
- Open XML SDK is a complicated but also powerful framework
- I can now forward code to a pro developer and **potentially** reduce our costs of developing the validators

Results II

- A wiki with information
- 2000+ lines of executable code
- A summer spent
- Prototyped in Oslo, Stockholm, Gothenburg, Copenhagen, Glasgow and Turin
- Prototyped in hotel rooms, on busses, on ferries, in airports, at the office, at home, late at night
- I kept my relationship

Why YOU should prototype

For work

- Demonstrate the feasibility of your analysis / proposed file format policy

At the personal level

- Coding is fun
- Learning to code is a source of confidence

Appendix slides

Proposed file format policy for XLSX Strict and ODS

1. No password or other read/write protection
2. Must comply with standard
3. Must have cell values
4. No data connections (keep cell values)
5. No external cell references (keep cell values)
6. No external object references
7. No RealTimeData (RTD) functions (keep cell values)
8. No printer settings
9. No absolute filepath to local directory

Other requirements under consideration

1. First sheet must be active
2. Embedded objects?
3. No macros

Tested CLISC prototype on 16.000 spreadsheets

Results

- COUNT: 16115 spreadsheets
- CONVERT: 658 of 16115 spreadsheets failed conversion
- ARCHIVE: 8 of 15457 converted spreadsheets have invalid file formats
- ARCHIVE: 15 of 15457 converted spreadsheets had no cell values
- ARCHIVE: 403 of 15457 converted spreadsheets had data connections - Data connections were removed
- ARCHIVE: 88 of 15457 converted spreadsheets had external cell references - External cell references were removed
- ARCHIVE: 0 of 15457 converted spreadsheets had RTD functions - RTD functions were removed
- ARCHIVE: 0 of 15457 converted spreadsheets had external object references - External object references were removed
- ARCHIVE: 11881 of 15457 converted spreadsheets had printer settings - Printer settings were removed
- ARCHIVE: 2527 of 15457 converted spreadsheets did not have active first sheet - Active sheet was changed
- ARCHIVE: 606 of 15457 converted spreadsheets have embedded objects - Embedded objects were NOT removed
- Total process time: 02:12:37:11 (days:hrs:min:sec)

Open XML SDK, Strict related issues

- [ ] Support validation of conformance OfficeDev/Open-XML-SDK#1223
- [ ] DocumentType Strict does not exist OfficeDev/Open-XML-SDK#1184
- [ ] Validation of Excel Strict files reports two errors OfficeDev/Open-XML-SDK#1190
- [ ] Strict Excel Files Have Invalid int32 Data Types OfficeDev/Open-XML-SDK#1142

Open XML SDK, other issues

- [ ] ChangeDocumentType (.xlsm) to (.xlsx) not removing /xl/vbaProject.bin OfficeDev/Open-XML-SDK#1209
- [ ] Absolute path is not found OfficeDev/Open-XML-SDK#1208

LibreOffice

- [ ] Converting .xlsx to .ods does not correctly convert chosen filtering options
- [ ] Support OOXML Strict Format in "Save as" GUI and "convert-to" CLI
- [ ] Support output filepath in CLI convert

Apache POI

- [ ] Support Strict OOXML files

Issues created at other repositories

Any questions, contact

Asbjørn Skødt, assk@sa.dk