Significant Properties Of Spreadsheets

An Update On The Work Of The Open Preservation Foundation's Archives Interest Group

Remco van Veenendaal
National Archives of the Netherlands, The Netherlands
Remco.van.Veenendaal@nationaalarchief.nl
https://orcid.org/0000-0002-2351-1677

Frederik Holmelund Kjærskov
Danish National Archives, Denmark
fhk@sa.dk

Kati Sein
National Archives of Estonia
Estonia
kati.sein@ra.ee

Jack O'Sullivan
Preservica
United Kingdom
jack.osullivan@preservica.com

Anders Bo Nielsen
Danish National Archives, Denmark
abn@sa.dk

Phillip Mike Tømmerholt
Danish National Archives, Denmark
pmt@sa.dk

Jacob Takema
National Archives of the Netherlands, The Netherlands
jacob.takema@nationaalarchief.nl

Extended abstract for poster proposal – In this extended abstract, the Open Preservation Foundation's Archives Interest Group reports on our ongoing investigation of significant properties of spreadsheets. Using the InSPECT methodology for investigating significant properties of electronic content, our goal is to get hands-on experience in investigating the significant properties of deposited spreadsheets by adding a Spreadsheet Testing Report to the InSPECT Testing Reports lore. An additional result of the AIG investigation is a Spreadsheet Complexity Analyser tool that extracts spreadsheet-specific properties and can be used to calculate the complexity of a spreadsheet based on the values of those properties.

I. Introduction

The Open Preservation Foundation's Archives Interest Group (AIG) started in July 2016. In order to establish the work priorities for the AIG, we used the elements of the SCAPE Catalogue of Preservation Policy Elements [1] as a long list of priorities. Each AIG member prioritised their copy of the list. Combining the lists resulted in our work priorities. One priority is an investigation of the significant properties of spreadsheets.

II. Why Investigate The Significant Properties Of Spreadsheets?

The AIG chose to investigate significant properties of spreadsheets, because (a) we wanted to get hands-on experience in investigating significant properties as a means of understanding the original deposited object, and how to preserve it and (b) as national archives, we receive more and more spreadsheets that are eligible for long-term preservation, but are faced with the current shortcomings of ensuring long-term accessibility of the spreadsheets while still preserving their significant properties. The Danish National Archives in particular had been asked to add suitable formats for preserving spreadsheets to their list of accepted formats, and in order to choose a format, needed to know which properties the format should be able to preserve.

The digital preservation community has

16th International Conference on Digital Preservation
iPRES 2019, Amsterdam, The Netherlands.

investigated significant properties in general and those of spreadsheets in particular, but there have been few significant properties of spreadsheets studies in recent years, while spreadsheet technology keeps changing. The few significant properties of spreadsheets resources available have e.g. been collected in the list of Significant Significant Properties [3], with 18 properties for spreadsheets, stemming from 2 resources. As AIG we found this too meagre a basis for decision-making and decided to start our own investigation of significant properties of spreadsheets.

III. Using The InSPECT Methodology

The AIG members created a recommended reading list about significant properties, collected spreadsheet example files and spreadsheet (file format) specification documentation as a knowledge base. We looked for significant property investigation methodologies and decided to use the InSPECT methodology for investigating significant properties of electronic content [2] for our investigation.

The InSPECT methodology is a well-documented formalized methodology that has been used and re-used in significant property investigations and resulted in a collection of Testing Reports[1]. We want to add our work to this lore.

IV. Object Analysis And Spreadsheet Complexity Analyser

The AIG followed the activities defined by the InSPECT methodology. At the time of writing, we are performing the Object analysis set of activities. We have selected spreadsheets as our object type, analysed the structure of spreadsheets by using (property extraction) tools and studying (file format) specifications and identified the purpose of spreadsheet properties by classifying them as one of the categories Content, Context, Appearance, Structure or Behaviour. We are currently working on steps 4, 5, and 6: linking behaviours to functions and structure.

One task of the InSPECT methodology is to get a list of tools that can be used to extract (technical) properties of electronic content. While listing and testing tools for extracting properties of spreadsheets (including Apache Tika[2], Dependency Discovery Tool[3] and the New-Zealand Metadata Extraction Tool[4]), we noticed that there were hardly any tools for extracting spreadsheet-specific properties, like used cells and worksheets, hyperlinks, formulas and scripts, embedded objects, pivot tables, etc.

Another challenge arose when we discussed possible subtypes of spreadsheets. Our initial thoughts were to have 'simple/static' spreadsheets vs. 'complex/dynamic' ones, where the former are mainly meant for pretty-printing tabular data on a single worksheet, and the latter for more complex calculations across more than one worksheet.

The combination of these two issues resulted in the need for a tool that can analyse the complexity of spreadsheets based on the values of extracted spreadsheet properties. This tool did not exist. We therefore developed a 'Spreadsheet Complexity Analyser', voted on which properties this analyser should be able to extract, and decided when a spreadsheet would be deemed 'simple/static' or 'complex/dynamic'. One test of the tool showed that 99% of a 180,000 Microsoft Office Excel file test set from the National Library of the Netherlands Ejournal collection were 'complex/dynamic'.

Even after revisiting our decision rules, we noticed that a categorisation in 'simple/static' and 'complex/dynamic' may be too simplistic. And that pre-programmed decision rules limit the possible uses of the tool. Users should e.g. be able to configure their own decision rules, as different organisations may have different decision criteria.

But even if we were to drop the idea of using the tool for distinguishing between different sub-types

[1] See e.g. https://web.archive.org/web/20160416031256/ http://www.significantproperties.org.uk/testingreports.html.

[2] Available from http://tika.apache.org/, accessed March 15, 2019.

[3] Available from https://sourceforge.net/projects/officeddt/, accessed March 15, 2019.

[4] Available from http://meta-extractor.sourceforge.net/, accessed March 15, 2019.

of spreadsheets, the tool seems to be a candidate for filling a gap in the property extraction and migration quality assessment tool ecosystem. If accepted, we will use the poster opportunity to discuss uses and improvements of the tool with our audience.

V. Conclusion

In this extended abstract, we presented the state of affairs of the AIG's ongoing investigation of significant properties of spreadsheets. We have chosen the InSPECT methodology for investigating significant properties of electronic content and are using that methodology to get hands-on experience in investigating the significant properties of spreadsheets. We are performing the Object analysis set of activities, and are in the process of linking spreadsheet behaviours to functions and structure. As a result of our work, a Spreadsheet Testing Report will be added to the InSPECT Testing Report lore.

Our preliminary conclusions from the Object analysis support earlier findings of significant property studies: the complexity and context-sensitivity of and degree of freedom inherent in spreadsheets makes creating an exhaustive list of significant spreadsheet properties practically impossible. But a list of (technical) significant properties does already help choose suitable file formats for preserving spreadsheet information. Further stakeholder analysis is required for fine-tuning our work.

An additional result of the AIG investigation is a Spreadsheet Complexity Analyser tool that extracts spreadsheet-specific properties and can be used to calculate the complexity of a spreadsheet based on the values of those properties.

Acknowledgements

The AIG would like to thank the OPF team: Becky McGuinness and Charlotte Armstrong for facilitating AIG meetings, and Carl Wilson for an introduction to GitHub and giving feedback on the Spreadsheet Complexity Analyser.

References

[1] Bechhofer, S., Sierman, B., Jones, C., Elstrøm, G., Kulovits, H., Becker, C.: Final version of policy specification model. http://www.scape-project.eu/deliverable/d13-2-catalogue-of-preservation-policy-elements (2014). Accessed March 15, 2019

[2] Knight, G.: InSPECT Framework Report. https://web.archive.org/web/20160520083956/http://www.significantproperties.org.uk/inspect-framework.html (2009). Accessed March 15, 2019

[3] Lucker, P., Sijtsma, C., van Veenendaal, R.: Significant Significant Properties. https://osf.io/rtjw3 (2018). Accessed March 15, 201