The Significant Properties Of Spreadsheets

Results From The Object Analysis And The Stakeholder Analysis By 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

Lotte Wijsman

National Archives of the Netherlands

The Netherlands

Lotte.Wijsman@nationaalarchief.nl

Asbjørn Skødt

Danish National Archives

Denmark

assk@sa.dk

Kati Sein

National Archives of Estonia

Estonia

kati.sein@ra.ee

Jacob Takema

National Archives of the Netherlands

The Netherlands

Jacob.Takema@nationaalarchief.nl

Jack O'Sullivan

Preservica

United Kingdom

jack.osullivan@preservica.com

17th International Conference on Digital Preservation

iPRES 2021, Beijing,China.

DOI: 10.1145/nnnnnnn.nnnnnnn

Abstract – When preserving spreadsheets, in certain cases it is decided to convert the spreadsheet file format to an image-based file format such as TIF. However, during this conversion, a loss of information can occur. Certain functionalities cannot be translated to the new file format and therefore lose their meaning. An example of this are formulas. When converting towards an image-based file format, the outcome of the formula will still be displayed, but the calculation that is behind it is lost. It is therefore imperative to find out which properties in spreadsheets are significant, such that these can be preserved. The Open Preservation Foundation's Archives Interest Group (OPF AIG) decided to research this using the InSPECT methodology, which consists of both an Object Analysis and a Stakeholder Analysis. These two types of analysis eventually led to a longer report which was published, in which not only our results were shared, but also a reusable and practice-tested method to apply the InSPECT methodology. In this paper, we present the main findings of our research.

Keywords – Significant properties, spreadsheets, preservation, Object Analysis, Stakeholder Analysis

Conference Topics - Building the Capacity & Capability; Enhancing the Collaboration

I. Introduction

The Open Preservation Foundation (OPF) is a global not-for-profit membership organization working to advance shared standards and solutions for the long-term preservation of digital content.[1] The OPF's Archives Interest Group (AIG) was formed in 2016 with the wish to do research and practical work together in order to mitigate challenges that all organizations in the group face.[2] One such challenge on every group member's mind was the preservation of spreadsheets. This type of information is highly dependent on the software environment in which it was created. Software – and the accompanying file format(s) - can become outdated and even obsolete. In the past this has happened with e.g. Lotus 1-2-3.[3]

When a file format becomes obsolete, one approach to preserve the information contained in the format is to convert[4] it to another file format – preferably a modern, open file format. Certain archives already convert their spreadsheets, or require their providers to convert their spreadsheets, because no spreadsheet file formats are considered acceptable.[5]

File format conversion has risks and consequences. Even if we assume that conversion software does a perfect job, not all conversions will preserve all properties of the original information type or the file format. The result of some conversions could be that the outcome of a formula could remain visible, but the underlying calculation (the formula itself) might not be preserved. By researching which properties are deemed significant in spreadsheets, archives and other cultural heritage organizations will be able to make better choices in deciding how to preserve spreadsheets.

To research the significant properties of spreadsheets, the AIG selected the InSPECT Framework Methodology.[1] This methodology's Requirements Analysis includes two main activities. The first step is to conduct an Object Analysis: "The evaluator analyses a representative sample of an object type, identifies a set of functions and behaviors that may be achieved, and the properties that are necessary for their performance." The outcomes of the Object Analysis are inputs for the second step, the Stakeholder Analysis: "The evaluator identifies one or more stakeholders that have some relationship with the Information Object and analyze the functions that they wish to perform." [1]

With the results of this research, an estimation can be made of how well conversion to another file format can preserve the information contained in a spreadsheet. From e.g. the work on Significant Significant Properties [2] it becomes clear that Object Analysis is a fairly well-established step in preservation research into significant properties. New in our research is that we have gained and hereby share

[1] https://openpreservation.org/.

[2] Currently, the members of the OPF AIG consists of individuals from the National Archives of the Netherlands, the National Archives of Estonia, the Danish National Archives, and Preservica.

[3] https://en.wikipedia.org/wiki/Lotus_1-2-3.

[4] We limit ourselves here to the preservation strategy conversion, because this is the strategy used by AIG members. Other preservation strategies may also involve significant properties.

[5] The Danish National Archives accepts a limited number of file formats for information to be transferred. These have not included spreadsheet specific formats in the past. Thanks to the AIG research, Denmark is working towards accepting spreadsheet file formats.

practical experience of three national archives with performing the Stakeholder Analysis. This not only resulted in recommendations with regard to significant properties of spreadsheets. Our experiences can also be reused by others to investigate the significant properties of other information types.

A. *Spreadsheets*

A spreadsheet is a file to organize, show, analyze, and manipulate data in tabular form. Data is stored in the table cells and can be either numeric, text, or results of formulae that calculate and display values based on the contents of other cells or an external data source.

Spreadsheet formats were created together with their main spreadsheet application, among them are VisiCalc, SuperCalc, Multiplan Lotus 1-2-3, Lotus Improv, Borland Quattro, Microsoft Excel, StarOffice, OpenOffice, and LibreOffice. Often several versions exist for each format (e.g. Excel 2010/2013/2016). Although it is possible to reuse spreadsheet formats among applications and application versions because there is a basic understanding between formats, this will in most cases result in a loss of information and/or functionality. The formats are originally tailored to the capabilities and operations of the (original) software applications, and why would one re-use formats in applications for which they are not originally intended. This explains why there is no comprehensive interoperability between spreadsheet formats and applications.

B. *Significant Properties*

By significant properties, we refer to the definition given by A. Wilson in his Significant Properties Report: “The characteristics of digital objects that must be preserved over time in order to ensure the continued accessibility, usability, and meaning of the objects.”[3]

Over the years, various terms have been used in the research for significant properties. Terms used were significant characteristics, significant properties, aspects, and essence.[4] However, this paper is not about (re)defining the term. We have decided to embrace the definition that is used the most by the international digital preservation community, namely significant properties.

C. *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.[1]

The InSPECT methodology is a well-documented formalized methodology that has been used and reused in significant property investigations and resulted in a collection of Testing Reports.[6] The AIG wants to add this research to this lore.

II. Object Analysis

A. *Methodology*

To conduct the Object Analysis, seven sub-tasks need to be followed according to the InSPECT methodology. By following these steps sequentially, the evaluator will increase their understanding of the technical composition and purpose for which the object type can be used. The seven sub-tasks are as following:

1. Select object type for analysis
2. Analyze structure
3. Identify the purpose of technical properties
4. Determine expected behaviors
5. Classify behaviors into functions
6. Associate properties with each function
7. Review and finalize

As a prerequisite, an evaluator must have the following resources to perform the Object Analysis stage:

- A representative sample of objects for analysis
- Technical specifications or standards that describe the composition of the object
- Characterization tools for analysis of the objects

The AIG selected spreadsheets as the object type, analyzed the structure of spreadsheets by using property extraction or characterization tools

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

and studying spreadsheet file format specifications. We identified the purpose of spreadsheet properties by classifying them as one of the categories Content, Context, Appearance, Structure or Behaviour:

- Content: characteristics of the content of the information object, such as the text, images, recorded sound, etc.
- Context: characteristics of the organizational, functional and operational environment at the time of creating, receiving, storing and/or using information objects, and any relations of the information object with other information.
- Appearance: features related to the visual presentation of information objects to certain users at the moment of interaction, such as font, size, layout, etc.
- Structure: characteristics that specify how parts of the information object are organized and related to each other, such as embedded objects, paging, headings, etc.
- Behaviour: characteristics that make behavior of, interaction with, or functionality of the information object possible, such as hyperlinks, formulas, etc.

Similar to what is explained in the InSPECT methodology, the work on the final sub-tasks was not a one-off process. It had many iterations.

As also explained in the InSPECT methodology, the Object Analysis and Stakeholder Analysis steps “may be performed in parallel or at different time periods.” In practice, we did both. Especially when preparing questionnaires for the Stakeholder Analysis, we noticed that it was sometimes useful to go back and make changes to the Object Analysis with regards to e.g. ordering or labelling to create more stakeholder-friendly questionnaires. In the Results section, we e.g. detail how we decided to group properties together using ‘industry standard’ property groups.

B. Results

In order to know which properties can be extracted from spreadsheets, we looked for and tried out these characterization tools[7]: FITS,[8] fido,[9] Siegfried,[10] Lingfo,[11] Dependency Discovery Tool,[12] Officeparser.py, [13] Ssconvert, [14] Python oletools, [15] Apache POIfs,[16] Apache Tika,[17] and (counted as one) several Python libraries to access spreadsheets.[18] The File Information Tool Set FITS is a toolset, and it includes some relevant tools for (extracting properties from) spreadsheets: Apache Tika (also investigated stand-alone), DROID, ExifTool, FFIdent, File utility, JHOVE, National Library of New Zealand Metadata Extractor, OIS File Information. We used all the listed tools on our test set of spreadsheets and obtained a long list of properties that could be extracted.

1. Spreadsheet Complexity Analyser

While listing and testing tools for extracting properties of spreadsheets, 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. Extracting these properties was necessary for us to divide spreadsheets into subtypes. 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. However, there was not yet a tool that could facilitate this.

The combination of these two issues resulted in the need for a tool that can analyze the complexity of spreadsheets based on the values of extracted spreadsheet properties. This tool did not exist. We

[7] Looking for characterization tools included using COPTR. https://coptr.digipres.org/index.php/Spreadsheet.

[8] https://projects.iq.harvard.edu/fits/home

[9] https://openpreservation.org/products/fido/

[10] https://github.com/richardlehane/siegfried

[11] http://web.archive.org/web/20190117194003/http://www.lexicon.net/sjmachin/xlrd.htm.

[12] https://sourceforge.net/projects/officeddt/

[13] https://github.com/unixfreak0037/officeparser

[14] https://www.systutorials.com/docs/linux/man/1-ssconvert/

[15] https://github.com/decalage2/oletools

[16] https://poi.apache.org/components/poifs/

[17] https://tika.apache.org/

[18] Mainly XLRD (https://pypi.org/project/xlrd/), pywin32 (https://github.com/mhammond/pywin32) and odfpy (https://pypi.python.org/pypi/odfpy).

therefore developed a 'Spreadsheet Complexity Analyser' (SCA).[5] The AIG reported on this development in a short paper on the updates of the work on the significant properties of spreadsheets. This was presented during iPRES 2019.[6]

This development is mentioned here because of the AIG's wish to be able to distinguish between 'simple/static' and 'complex/dynamic' spreadsheets.[19] For the former, a file format that can't preserve dynamic spreadsheet properties might still be suitable (e.g. TIFF or PDF). And in e.g. the case of the Danish National Archives, they already have a lot of knowledge of and experience with these formats. Being able to continue to use these formats for spreadsheets might be preferable from a business perspective. For the latter, a spreadsheet-specific file format would be required (e.g. ODS or XLSX). Using a tool like the SCA, one can measure the number of both subtypes of spreadsheet and decide which spreadsheet to convert to which file format. Business rules could help decide when an organization has enough spreadsheets of the second type to start acquiring more knowledge about spreadsheet-specific file formats and (conversion) tools.

The AIG collected technical specifications of spreadsheet file formats. Several spreadsheet formats exist. They were created together with their main spreadsheet application. Among them are VisiCalc, SuperCalc, Multiplan, Lotus 1-2-3, Lotus Improv, Borland Quattro, Microsoft Excel, Open Office and Libre Office.[7] Often several versions exist for each format. We went over the specifications and collected the properties.

The combination and de-duplication of the properties extracted by tools and the properties extracted from the specifications resulted in a list of 200 spreadsheet properties, stored as a spreadsheet worksheet on our shared drive.

2. Identify Purpose of Technical Properties

Each of the 200 properties were manually classified as one of the categories Content, Context, Appearance, Structure or Behaviour. Other information was added to this knowledge base, such as its origin (which specification or tool). As the column headers of this knowledge base were in blue, this list became known as the 'blue sheet'.

As we started working towards connecting the properties to purpose, behavior and function, property group names that reflected those characteristics were introduced: e.g. Security for any spreadsheet security-related properties and Character Formatting for character and cell formatting properties. We noticed that having a vast amount of properties and ad hoc property group names would make talking to stakeholders about significant properties difficult, which is why Frederik Holmelund Kjærskov of the Danish National Archives proposed to use 'industry standard' compatibility table property groups used by e.g. Microsoft and Apple.[7] As most of the spreadsheets the national archives in the AIG receive are of a fairly recent date – less than 20 years in most cases - and stakeholders will nowadays not be likely to talk about Lotus 1-2-3 spreadsheets, we felt that it was warranted to use these tables.

3. Determine Expected Behaviors, Classify Behaviors Into Functions, and Associate Properties With Each Function

We then fed the property group information from the blue sheet into a new work sheet and added expected behaviors, classified those behaviors into functions and associated the property groups with these functions. This sheet became known as the 'green sheet'.

4. Final Report

More details on the Object Analysis step, together with the underlying data were published were published in the final report of the AIG's investigation of significant properties of spreadsheets. [7]

III. Stakeholder Analysis

A. Case Studies

[19] The SCA has default values to recognize if a spreadsheet is considered simple/static or complex/dynamic. See https://github.com/RvanVeenendaal/Spreadsheet-Complexity-Analyser/blob/master/SpreadsheetComplexityAnalyser.config.

For the Stakeholder Analysis, there was a deviation from the InSPECT methodology. It was felt that the methodology was slightly abstract and could therefore be difficult to implement in interviews with stakeholders. Deviating from the methodology also allowed us to use more diverse ways to perform stakeholder analyses. This would allow us to learn from each other which approaches were successful, and which were less successful, and come with more extensive recommendations. Furthermore, all three of the National Archives in the AIG had different aims for their case studies. This diversity would enrich the research more by having different views and perspectives.

1. National Archives of the Netherlands (NANETH)[20]

The aim of NANETH's Stakeholder Analysis was to discover whether any property groups are deemed significant by stakeholders. To this end, individuals employed in the public sector were sought out. This is due to the National Archives of the Netherlands primarily preserving information from public institutes. By selecting accompanying stakeholders, it is ensured that the spreadsheets considered in this research are similar to the spreadsheets that would eventually be preserved by members of the OPF AIG.

Further factors that should be considered when selecting participants are diversity in their knowledge regarding spreadsheets and the role they have in relation to spreadsheets. Asserting different roles is in accordance with the InSPECT methodology, which states that there should be a clear understanding regarding the relationship of the stakeholder with the relevant information object. Having representatives of every group in the sample could help to ascertain if there were differences in what each group deemed significant. Three roles were therefore distinguished: maker, user, and manager. It, however, later became evident that the roles of maker and user are often intertwined. 16 stakeholders were eventually found, of which seven were employed by Dutch ministries, whilst the other nine work for semi-governmental institutes. Furthermore, the participants were diverse in their level of knowledge and their roles.

The first step of the analysis was to gain insight in the background of the stakeholder through preliminary questions. For instance, their level of knowledge, type of role, and which properties they found to be significant were queried. Additionally, stakeholders were asked to submit a spreadsheet representative of their daily use. This spreadsheet may help explain some of their choices, for instance formulas would be an unlikely choice for participants that do not use them. The SCA was used to extract information on properties used in the representative spreadsheets.

Subsequently, more targeted questions were sked. The list of properties created during the Object Analysis was used for this, and individual properties were subsumed into groups. A total of 21 groups were created. Examples of these groups are cell formatting and formulas. A catalogue was created to further clarify the groups.[21] This catalogue was sent to participants, asking them to pick and motivate the five groups they deemed most significant. The final step of the analysis was then to conduct semi-structured interviews, to obtain further clarifications where needed.

Results from the catalogue showed five distinct groups that were deemed significant:

- Formulas (chosen 13 times). A formula calculates the value of a cell (or multiple cells).
- External Data (chosen 10 times). This is data that exists outside of the application itself. External data is retrieved from an external source via queries and may be dynamic.
- Cell Content (chosen 9 times). Cell content is (in this study) any text stored in cells.
- Pivot Tables (chosen 8 times). A pivot table is a table that summarises data of a more extensive table into key statistics, such as the mean and sums.
- Charts (chosen 7 times). A chart visually displays data in various ways, such as bars or a pie.

[20] A more extensive report on the Stakeholder Analysis conducted by the National Archives of the Netherlands can be found at http://doi.org/10.5281/zenodo.3971833.

[21] L. Wijsman. (2020, June 20). Catalogue Significant Properties of Spreadsheets. Zenodo. http://doi.org/10.5281/zenodo.3902080.

It should be noted that out of the three stakeholders that did not select formulas, one stated that this was due to never making use of them. This shows that the choices made are highly dependent on factors such as background and level of knowledge. Four of the five groups that were most selected are dynamic groups. This underlines the notion that converting spreadsheets to image-based file formats (e.g. TIF) is insufficient.

Concluding, NANETH's Stakeholder Analysis confirmed our choice to preserve spreadsheets in original file formats. Converting spreadsheets to image-based file formats does not meet the needs of the stakeholders, who often deem dynamic properties to be significant.

2. National Archives of Estonia (NAE)

The objectives for the stakeholder interview were to (1) figure out what aspects of spreadsheets the producers of spreadsheets regard as important themselves, (2) whether colors, fonts, styles, or other formatting is used to carry semantic information, and (3) to get a very rough understanding of the proportion of simple/static vs complex/dynamic type of spreadsheets created.

As a national archive, the obvious choice was to contact document managers at public sector agencies that were obliged to sooner or later submit spreadsheets to the archive. This way we would learn only about the files that will eventually be preserved by the NAE. Moreover, we wanted to get in contact with stakeholders that were producers of spreadsheets. The closest producer found was the NAE itself. Eventually, interviews were held with a document manager (archivist) and an IT support employee of the NAE.

To investigate the objectives mentioned earlier, observation and oral interviews were selected. These were chosen intentionally to detect any unwritten rules or 'inner folklore' regarding to the creation and usage of spreadsheets. The stakeholders were contacted via a contact person within the NAE. The stakeholders were sent a sample questionnaire previously developed by the Danish National Archives (DNA) in advance.[7] The interview was held with both stakeholders together. During the interview, some spreadsheets that are of archival value were also shown. During the interview, the questionnaire was not followed strictly to keep the conversation flowing, but to also follow the stakeholders when something stood out as significant.

Resulting from the interview and observation, several findings became evident. When conducting a Stakeholder Analysis about the significant properties of spreadsheets, a spreadsheet producer may simply reply that in their spreadsheet, 'all properties are important'. Therefore, it is important to keep digging and use examples in the interview. Furthermore, it is crucial to make sure that both the interviewer and the interviewee understand the terms used in the same way. It is also important to understand who the producer is to find meaning behind certain properties used. Text and cell formatting, coloring included, is widely used in spreadsheets to give meaning to certain cells. However, this meaning is often not documented anywhere. Capturing this 'inner folklore' by observing and interviewing the producer of the spreadsheet is often a good method to use. Both static and dynamic types of spreadsheets were also under observation. The dynamic type was often used for budgeting. Unfortunately, there were too few stakeholders in this research to make a generalization here.

Overall, a Stakeholder Analysis is a beneficial way to learn about the creation and usage of the born digital content you wish to preserve in the future. Probably the best results are when the preservationist does conduct the Stakeholder Analysis together with the archivist and/or future curator of the collection to get a complete overview, including the 'inner folklore'.

3. Danish National Archives (DNA)

The DNA receives hundreds of thousands of TIF-converted spreadsheets yearly from public authorities. Despite TIF-converters being increasingly more refined, we still regularly encounter errors during validation and we are aware of the possible loss of significant properties, which imaging of spreadsheets entail. The DNA wanted to investigate these problems deeper and methodically. To do this we had to develop new methods, interview stakeholders, pilot test tools, and pool expertise on the subject through international cooperation such as the OPF AIG.

We developed targeted stakeholder questionnaires to use in physical interviews with two public data producers and two archives. We pilot tested available validators and we also applied the Spreadsheets Complexity Analyser tool to understand the distribution of simple versus complex spreadsheets and applied the same tool for characterizing data in our testbed for conversion experiments. This was done to test the loss of significant properties (and general abnormalities) when converting between Excel and OpenDocument Spreadsheet. Some conversion errors were encountered.

The interviews provided us with valuable insight in the different use cases of spreadsheets and made it possible for us, in varying degrees, to map significant properties through their eyes and experience. We saw that some content can be preserved through imaging, but by doing so the underlying structures are lost for good, and these structures are sometimes the only option to document the origin and interpret the content. This is for instance the case with formulas, references to defined names, the data areas of graphs, conditional formatting, and calculated values for pivot charts. Furthermore, by imaging, structures necessary for users' future interaction and navigation, such as sorting and filtering the spreadsheet, are lost. Especially for large spreadsheets this loss is unacceptable for users, because the practical limitations in navigating, reading, and understanding hundreds of printed pages is in sharp contrast to the preservation objective of being able to reuse data.

In conclusion, the stakeholders identified structures, which we currently do not preserve by converting to TIF, as significant to their use of spreadsheets. This had led us to consider adopting a dedicated spreadsheet format as the new accepted preservation format for this content information type.

B. Results

The case studies carried out by the three National Archives tried to not only establish which properties were deemed to be significant, but also how to perform a proper Stakeholder Analysis and if current practices are sufficient to the needs of stakeholders.

An important finding by all three National Archives was that dynamic content was deemed to be of significance. Formulas, external data, and pivot tables were chosen to be of the most significance by the stakeholders questioned by the NANETH. However, it is important to note here that what is deemed to be significant is highly dependent on what the spreadsheets in question contain. If a spreadsheet does not contain any formulas, the stakeholder is not likely to find these significant. This is why tools like the SCA may be useful in the preparatory phases of any Stakeholder Analysis.

Since the InSPECT methodology concerning the Stakeholder Analysis was very abstract, we felt that it would be difficult to use this in practice with stakeholders. Therefore, every National Archive developed its own method, which resulted in various results and lessons learned. These lessons could be applied to future stakeholder analyses. The primary finding was that there needs to be a clear understanding between the interviewer and the stakeholder concerning used terminology. As was the case with records vs (technical) files: the fact that one record may consist of several spreadsheet files starts to matter when you ask stakeholders about quantities. Come to a common understanding of these terms in the beginning and return to it a couple of times throughout the interview (for example, by saying "…and this information is about records, not files, right?").

It is important to be prepared when performing interviews. Before the interview, try to get an understanding of the content, in this case spreadsheets, that are of archival value in the organization. Bring example files, or ask the stakeholder to share examples before the interview. Even if the content of those specific spreadsheets is not relevant to the broader goal of the interview, the stakeholder will talk about and refer to these. This may be the basis of the categorization of different types of spreadsheets for the stakeholder (e.g. budgeting, reports etc.). Perhaps only certain types have certain properties that are deemed to be significant. This also works into understanding the stakeholder and their background. Knowing what they encounter daily can help to understand not only what they deem to be significant, but also why this is. Language also works into this by making sure that both parties understand core terms the same way. This was also why we used the compatibility tables by Microsoft and Apple: to connect to the stakeholders experience.

After assessment of the three stakeholder analyses that were carried out, we found it evident that some current practices are not sustainable. Converting spreadsheets to image-based file formats is no longer a viable approach when dealing with dynamic content such as formulas and pivot tables. For the Danish National Archives this also meant a possible revision of their accepted preservation formats. Adopting spreadsheet-specific formats such as XLSX and/or ODS could solve the problems that are currently encountered.

IV. Combining Object And Stakeholder Analysis Results

By combining results from the Object Analysis and Stakeholder Analysis, we were able to establish which property groups and properties are seen as significant by the stakeholders. We had also labelled property groups and property groups ourselves, which enabled us to calculate how well we as archive stakeholders were able to 'predict' the results of the Stakeholder Analysis. This combination led to two insights:

At the property level, we predicted 32% of the properties that were deemed significant in the Stakeholder Analyses. In 36% of the cases, the properties deemed significant in the Stakeholder Analysis matched our prediction. Looking at the same analysis at the property group level, the percentages were 49% and 94%.

While the significant property groups of spreadsheets were the same, the differences between what we as archive stakeholder predicted as significant properties and the results of the Stakeholder Analyses resulted in a 'long list' and a 'short list' of significant properties of spreadsheets. There were too many uncertainties in our results to claim that we had found 'the' significant properties of spreadsheets. When investigating the significant properties of your spreadsheets, start with the short list, add (what you think are) relevant properties from the long list and talk to your stakeholders to establish your final list of significant properties of spreadsheets. The lists are available in our final report. The significant property groups of spreadsheets are: Application Settings, Cell Content, Cell Formatting, Charts, Data Tools, Editing, External Data, Formatting, Formulas, Graphic Elements, Metadata and Pivot Tables.

V. Conclusion

In order to find the properties during the Object Analysis, two consecutive steps were taken in our research. The first step concerned tools. At the start, several characterization tools were used to identify which properties were present in spreadsheets. These tools are mostly capable of extracting properties at surface level and focus predominantly on file properties that can be seen in the spreadsheet application by the user. Therefore, for the purpose of the Object Analysis, which strives towards an in-depth overview of all properties present in spreadsheets, the characterization tools were deemed to be insufficient. Hence, our research led us to the creation of the SCA, which formed an addition to the other tools by extracting information about cells, sheets, formulas, named objects, macros, etc.

After using the characterization tools, in the second step of the Object Analysis our research focused on the specifications of different spreadsheet formats. Various spreadsheet formats can have distinctive compositions that are specific to a certain spreadsheet format and can therefore contain different properties. However, we found that these specifications are focused on the internal and more technical build-up of the format and are difficult to link to the actual use and function. This led us to also look at the compatibility tables between Open Document Format, Microsoft Excel and Apple Numbers. The compatibility tables are more suited for identifying properties that are linked to use and functionality. Moreover, they are compliant with terminology used by spreadsheet users in real-life.

Overall, it is important to keep the Stakeholder Analysis in mind when conducting the Object Analysis. The addition of several steps such as grouping and the compatibility tables of Microsoft and Apple ensures a smooth transition between the two types of analysis while also allowing you to interview stakeholders that have different levels of knowledge.

We were able to establish the significant property groups of spreadsheets but found that there were too many uncertainties at the individual property level to claim that we found the significant properties of spreadsheets. We therefore settled for a short list and a long list of significant properties. When working on significant properties of (your) spreadsheets, start with the short list and add (what you think are) relevant properties from the long list to create the baseline for stakeholder interviews.

The most important conclusion is related to this: our results demonstrate that performing a Stakeholder Analysis is important. As archive stakeholders, we can only predict one third (at the individual property level) to half (at the property group level) of the properties that other stakeholders deem significant.

More details surrounding the Object Analysis and the Stakeholder Analysis, with the underlying data, are published in the final report of the AIG's investigation of significant properties of spreadsheets.[7]

Our recommendations for future research are to further refine the research done with the significant properties of spreadsheets. We got the impression from talking to colleagues that few Stakeholder Analyses have ever been performed, and fewer using the InSPECT methodology. As a community, we need to get more experienced performing these analyses.

While there is no 'one size fits all' solution, categories could be created that fit a certain type of spreadsheet that has a particular set of significant properties. Following this, research will need to be done to assign fitting formats to preserve the significant properties concerning that category. For instance, when a spreadsheet is mostly static and does not make use of formulas or pivot tables, converting it to a TIFF could be acceptable. However, if formulas are used in a spreadsheet and are considered to be significant, another format must be selected that optimizes the preservation of this category of spreadsheet.

Concluding, our research into the significant properties of spreadsheet resulted in a list of significant property groups and a short and long list of significant properties. It gave us several re-usable practices and yielded valuable insights. Conducting the Object Analysis resulted into an extensive list of spreadsheet properties. These properties were then connected to behaviors and functions. Furthermore, the SCA was created. It is an open-source tool that others can use to analyze their own spreadsheets or to prepare for the Stakeholder Analysis. Additionally, the Stakeholder Analysis resulted into three different reusable practices that can be used by the preservation community. Overall, our results and final report have not only brought us results, but also contributed towards a reusable and practice-tested method for applying the InSPECT methodology.

ACKNOWLEDGMENT

The AIG would like to thank the OPF team: Becky McGuinness and Charlotte Armstrong for facilitating the monthly online AIG meetings, and Carl Wilson for an introduction to GitHub and giving feedback on the SCA. We would also like to thank our Danish colleagues Frederik Holmelund Kjærskov, Anders Bo Nielsen, and Philip Tømmerholt for their contributions in earlier stages of this research.

REFERENCES

[1] Knight, G., "InSPECT Framework Report." https://web.archive.org/web/20160520083956/http://www.significantproperties.org.uk/inspect-framework.html (2009).

[2] Van Veenendaal, R, P.C.M. Lucker & C.D. Sijtsma. "Significant significant properties." https://openpreservation.org/wp-content/uploads/2018/10/Significant-Significant-Properties.pdf

[3] Wilson, A. "Significant Properties Report." https://significantproperties.kdl.kcl.ac.uk/methodology.html

[4] Dappert, A., and A. Farquhar, Significance Is in the Eye of the Stakeholder, Proceedings of the 13th European conference on Research and advanced technology for digital libraries (EDCL 2009): p. 298.

[5] Github. "Spreadsheet Complexity Analyser." https://github.com/RvanVeenendaal/Spreadsheet-Complexity-Analyser

[6] Van Veenendaal, R. et al. "Significant Properties of Spreadsheets: An update on the work of the Open Preservation Foundation's Archives Interest Group." iPres 2019: 396 - 398.

[7] The Significant Properties of Spreadsheets (OPF AIG Final Report). https://doi.org/10.5281/zenodo.5387099.