The Significant Properties of Spreadsheets

A report by the Open Preservation Foundation's Archives Interest Group

Version 1.0

Open Preservation Foundation

Table of Contents

1. Introduction

1.1. Who is the OPF Archives Interest Group?

The Open Preservation Foundation (OPF) brings together institutions with the common goal to do digital preservation. The Archives' Interest Group (AIG) was formed by OPF members in 2016 to carry out research and practical work collaboratively in order to mitigate one real challenge that everybody in the group faces. This report is the end result of the research on the significant properties of spreadsheets. The current members of the AIG are from the National Archives of the Netherlands (NANETH), the National Archives of Estonia (NAE), the Danish National Archives (DNA), and Preservica. A full list of all AIG colleagues who contributed to this work is listed in Appendix D.

1.2. Why are we doing this?

Preserving files in spreadsheet formats is a priority for every member. We need to answer questions such as 'should we migrate?' and 'how do we measure the success or quality of the migration?'. For the latter, we need to know what aspects of the file are important (significant), which led us to the decision to investigate the significant properties of spreadsheets.

1.3. What is a spreadsheet

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 are 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 re-use 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.

1.4. What are significant properties

By 'significant properties', we refer to the definition given in Andrew Wilson's 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, and their capacity to be accepted as evidence of what they purport to record."[1] When the digital objects (e.g. files in a format) or the technology to use them (e.g. viewers) are at risk of becoming obsolete, preservation actions may be required (e.g. file format migration or viewer software emulation). Ensuring that the significant properties are reasonably preserved as a result of these preservation actions is then a means of validating these actions.

[1] A. Wilson, "Significant Properties Report," https://significantproperties.kdl.kcl.ac.uk/wp22_significant_properties.pdf, p. 8.

Over the years, various terms have been used in the research of significant properties, including significant characteristics, significant properties, aspects, and essence.[2] However, this report will not (re)define the term. We have decided to embrace the definition that is used most frequently by the international digital preservation community: significant properties.[3]

1.5. Preservation challenges

Preservation challenges of a spreadsheet occur when:

- Opening the file in a software application other than the original creating software;
- Determining the purpose of the information object (e.g. via analysing the values of properties);
- Migrating from one spreadsheet format to another, or to a non-spreadsheet format.

Data loss and loss of functionality are likely to be the most common problems.

In most cases, it's not always clear what the capabilities are between file formats and software products used to render them. However, sometimes you are lucky, when the capabilities of software products are already declared by vendors, as is the case of Microsoft Excel rendering Open Document Format.[4]

We carried out a test to observe the treatment of decimal places. Inspired by an earlier OPF blog post,[5] we experimented by rendering and converting an XLS file with different software in order to ascertain the issue with the number of decimal places calculated. In particular, we observed how a cell that contained the formula of the type AVERAGE was rendered in Microsoft Excel, Open Office, and LibreOffice. Each of these products calculated the average with a slightly different result, illustrated in the second column of the table below. Each software application calculates a different number of decimal places by default, either 9, 10 or 14.

Renderer	Value in XLS	No. of decimal places	Value in CSV	Value in ODS
OpenOffice 4.1.3 Calc	9.5963934426	10	9.5963934426	9.5963934426
LibreOffice 5.2.4.2 Calc	9.59639344262295	14	9.59639344262295	9.59639344262295

[2] A. Dappert & 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.

[3] P. Lucker, C. Sijtsma & R. Van Veenendaal, "Significant Significant Properties – Award Winner: Popular Poster," June 20th, 2019: https://osf.io/rtjw3/.

[4] "Differences between the OpenDocument Spreadsheet (.ods) format and the Excel for Windows (.xlsx) format," Microsoft, https://support.microsoft.com/en-us/office/differences-between-the-opendocument-spreadsheet-ods-format-and-the-excel-for-windows-xlsx-format-3db958c8-e0ac-49a5-9965-2c2f8afbd960?ui=en-us&rs=en-us&ad=us.

[5] J. Van der Knijff, "PDF/A as a preferred, sustainable format for spreadsheets?" OPF blog, December 9th, 2016, https://openpreservation.org/blogs/pdfa-as-a-preferred-sustainable-format-for-spreadsheets/.

MS Excel 2010	9,596393443[6]	9	9.596393443	Not available

Exact values of one cell in the observed XLS file and its derivatives, rendered with three different software programmes.

We converted the spreadsheet into CSV format, using the same rendering software respectively. The average calculated was always the same that the individual software showed during rendering (see 4th column in the table). An interesting side-observation was that CSV files opened in the Google Drive application showed even higher rounded up values.

Migration to ODS format was done with Excel 2016. The formula was preserved and thus the calculated value depended on the rendering software again (see the fifth column of the above table).

In conclusion, if an archive were to ingest the above-mentioned file into a digital repository, decisions should be made by the archivists and the provider: what is more important to preserve here - the formula or the value? If the value is more significant to preserve, the archivist should be extra careful to document the name and version of the original software (and its default settings) that created the file, or any human input about the expected precision.

Some organisations use a normalisation strategy to convert spreadsheets to archival formats such as PDF/A or text. For complex spreadsheets[7] this can lead to data loss during normalisation. In another example we compared an original MS Excel file with its normalised representation in PDF/A format and encountered some examples of data loss:

- Formulae were lost. There was no hint that there had been a formula at all or what ranges of data had been used for calculation.
- Names of worksheets were lost.
- Data of one worksheet was divided between different pages in the PDF. It was split both horizontally as well as vertically and the original location of data was unclear.
- It was only vaguely understandable where data from one worksheet ended and another began.
- Notes were not included in PDF.
- Excel-like references to a cell (e.g. E18) were not possible any more. The PDF file contained neither row numbers nor column headings making it almost impossible to refer to a particular cell.
- Unable to detect hidden rows and columns.

[6] In Excel, when used the "Enhance decimal" button one could go up to 14 digits. After this 0's were suffixed: 9,5963934426229500000. “Numeric precision in Microsoft Excel,” Wikipedia, https://en.wikipedia.org/wiki/Numeric_precision_in_Microsoft_Excel.

[7] See 3.1.1 for more information on the divide between ‘complex’ and ‘simple’ spreadsheets.

2. Previous Work

During the course of this research, the AIG has already reported back to the digital preservation community twice during iPRES with short papers and posters. In 2018, the National Archives of the Netherlands presented a related short paper and poster on 'Significant Significant Properties'.[8] An update on our work was subsequently given during iPRES 2019.[9] The work discussed in these short papers and posters will also be given in this report.

In our own previous work, we included previous significant properties work of the giants on whose shoulders we stand. In this report, we decided to focus only on previous work that was of direct relevance for our investigation: the work done in the JISC-funded Investigating the Significant Properties of Electronic Content Over Time (InSPECT) project[10] and especially the InSPECT Framework Report.[11]

2.1. Looking for the proper methodology

The starting point for the current study on spreadsheets was to carry out a literature review to build upon the work already done in this area, rather than reinvent the wheel. By gathering conference papers, project reports, government guidelines and other resources into a shared reading list helped the group have access to the same level of knowledge in preparation for the work ahead.

The analysis of significant properties is a well-established and recognized approach within the digital preservation community. Previous frameworks that use this kind of analysis in various degrees are Rothenberg & Bikson (1999),[12] the CEDARS project,[13] RLG,[14] Digital Preservation Testbed,[15] and DELOS.[16] Contemporary to the time of Knight's formulation of the InSPECT framework are the frameworks part of CASPAR[17] and PLANETS.[18] Since Knight's formulation, no other major frameworks have been published.

---

[8] P. Lucker, C. Sijtsma & R. Van Veenendaal, "Significant Significant Properties."

[9] R. Van Veenendaal et al, "Significant Properties of Spreadsheets," https://doi.org/10.17605/OSF.IO/G8D5Y.

[10] "Significant properties and digital preservation," Significant Properties (archived version), https://web.archive.org/web/20160520082501/http:/www.significantproperties.org.uk/.

[11] "InSPECT Framework Report," Significant Properties (archived version), https://web.archive.org/web/20160520083956/http://www.significantproperties.org.uk/inspect-framework.html.

[12] J. Rothenberg & T. Bikson, "Carrying Authentic, Understandable and Usable Digital Records Through Time," (Santa Monica, CA: RAND Corporation, 1999), https://www.rand.org/pubs/rand_europe/RE99-016.html.

[13] M. Jones, "The Cedars Project," Library and Information Research 26, no. 84 (2002). http://dx.doi.org/10.29173/lirg136.

[14] "Research Libraries Group," Wikipedia, https://en.wikipedia.org/wiki/Research_Libraries_Group.

[15] R. Verdegem & J. Slats, "Practical experiences of the Dutch digital preservation test-bed," VINE 34, no. 2 (2004): 56-65. https://doi.org/10.1108/03055720410531004.

[16] S. Strodl et al., "The DELOS Testbed for Choosing a Digital Preservation Strategy," Springer, http://dx.doi.org/10.1007/11931584_35.

[17] CASPAR Preserves, http://casparpreserves.digitalpreserve.info/.

[18] PLANETS, https://planets-project.eu/.

The aforementioned frameworks served, according to Knight, as useful inspirations that qualified the InSPECT framework but he also saw these as insufficient. We agree with this notion and have found that, in particular, the ability to tie significant properties with stakeholder analysis is a crucial advantage of this framework. Another important aspect is the use of the engineering design method: Functions, Structures, and Behaviours (FSB), which enabled us to apply common classification terminology to significant properties of data objects.

The AIG was unable to find any other groups actively working in the area of significant property stakeholder analysis. Most of the existing research into significant properties – and especially work on stakeholder analysis methodologies – was from 2009 or earlier, but still provided a useful exercise to identify previous approaches. We decided to adapt the InSPECT Framework Report as a structure for our research.

2.2. What is the InSPECT framework?

The InSPECT Framework Report (Investigating the Significant Properties of Electronic Content Over Time) was written by Gareth Knight in 2008 and updated to version 2.0 the year after. Knight was, at the time, employed at The Centre for e-Research at King's College in London. He collaborated with The National Archives to develop and write the method, which was funded through JISC.

The InSPECT Framework Report provides a methodology on how to execute two types of analyses: an Object Analysis and a Stakeholder Analysis. As mentioned in our previous paper on significant properties,[19] the InSPECT methodology is a well-documented formalized methodology that illustrates how to investigate the significant properties of electronic content. Several test reports of electronic content already exist (e.g. raster images and e-mail) and we would like to add spreadsheets to this set.[20]

In the rest of this document, "InSPECT methodology" is short for "the methodology detailed in the InSPECT Framework Report."

[19] R. van Veenendaal et al., "Significant Properties of Spreadsheets: An Update on the work of the Open Preservation Foundation's Archives Interest Group," iPRES 2019 - 16th International Conference on Digital Preservation.

[20] "Testing Reports," Significant Properties (archived version), https://web.archive.org/web/20160416031256/http://www.significantproperties.org.uk/testingreports.html.

3. Methodology, Tools, and Data

The InSPECT methodology equipped in this research consists of two types of analysis: the Object Analysis and the Stakeholder Analysis. In addition to expounding on this method, there will also be information on the tools, data, and the case studies conducted by the three National Archives taking part in AIG.

The AIG met in virtual monthly meetings and yearly physical meetings. In the meetings, work done was discussed, new work was started or continued collaboratively, and actions were set for the next meeting. The OPF team kept notes and provided an email list.

This chapter explains how we applied the InSPECT methodology in our work. The Results chapter presents the results of that work.

3.1. Object Analysis

When following the InSPECT methodology, seven sub-tasks need to be followed to conduct the Object Analysis to increase an 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. Analyse structure
3. Identify the purpose of technical properties
4. Determine expected behaviours
5. Classify behaviours into functions
6. Associate properties with each function
7. Review & finalise

How we applied the seven sub-tasks is explained in the next paragraphs.

3.1.1. Select object type for analysis

For this research, the AIG chose to investigate the significant properties of the “high-level object type” spreadsheets. With “high-level”, we mean to indicate spreadsheets in general and not specific file formats such as Microsoft Excel spreadsheets or Open Document Spreadsheets.

As national archives we receive an increasing amount of spreadsheets that are eligible for long-term preservation, but we are faced with the current shortcomings of ensuring the long-term accessibility of the spreadsheets while still preserving their significant properties. The Danish National Archives (DNA) 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. Subsequently, this research will yield hands-on experience on how significant properties should be investigated as a means of understanding the original deposited object, and how to preserve the object type while maintaining accessibility.

According to the InSPECT framework methodology, the evaluator must possess the following to perform the object analysis stage:

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

Representative sample
In addition to collecting publicly available spreadsheets and corpora with spreadsheets in them, the KB (where one of the authors, Jacob Takema, worked at that time) and DNA analysed thousands of non-public spreadsheets from their repository and made their findings available to the AIG.

Technical specifications
The specifications of spreadsheet file formats give insights on the components of which a file format is constructed. Specifications are also helpful to identify the properties of file formats. As explained above, the AIG wanted to investigate spreadsheets, not file formats. But by comparing property lists from technical specifications, we were able to abstract from specific spreadsheet file format properties to more general spreadsheet properties. When multiple spreadsheet formats have properties for 'cell', 'worksheet', 'formula', 'hyperlink', etc. we felt that it was safe to assume that these were generic spreadsheet properties.

Characterisation tools
"Characterization is the process of extracting specific characteristics from the file."[21] In order to know which properties can be extracted from spreadsheets, we found and tested these characterisation tools: FITS, fido, Siegfried, Lingfo (XLRD), Dependency Discovery Tool, Officeparser.py, Ssconvert, Python oletools, Apache POIfs, Apache Tika and (counted as one) some Python libraries to access spreadsheets.[22] 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 these on our test set of spreadsheets and obtained a long list of properties that could be extracted.

Sub-types
The InSPECT methodology gives the evaluator the option "to select a high-level object type (raster images, audio recordings, web pages, e-mail) or a sub-type that contains specific characteristics". As AIG, our main focus was the high-level object type spreadsheets. But being able to migrate spreadsheets to suitable file formats that preserve the spreadsheet's significant properties best, was a practical use case within the broader investigation. That is why we also included sub-types in our work.

Spreadsheets are often used for simply presenting information in tabular form (e.g. agenda or list) and not for complex calculations. Such simple spreadsheets would likely render more or less the same in most spreadsheet-rendering applications at every moment of time. One would lose no information when migrating to other, primarily rendering-oriented file formats, like the Tagged Image File Format (TIFF) currently accepted by DNA. On the other hand, there are more complex spreadsheets in which values in a cell are dependent on current date, values of other cells in the same spreadsheet or even

[21] L. Shala & A. Shala, "File Formats – Characterization and Validation," IFAC-PapersOnLine 49, no. 29 (2016): 253-258. https://doi.org/10.1016/j.ifacol.2016.11.062.
[22] See Appendix A for more information on characterisation tools used.

external sources, or that contain graphs or pivot tables. When rendering or migrating such complex spreadsheets, it is more likely that information may get lost. Extra checks are required.

As AIG, we defined 'simple/static' spreadsheets as spreadsheets that are mainly used for (human) visualisation and contain static data values organised into tabular format on one or more worksheets. The data values are not meant to change, and if they do, no other data values change. External information, like a different date or an updated external data source, do not result in changes in the spreadsheet. No significant loss of information would occur if these spreadsheets were migrated to Comma Separated Values files. In case of 'simple/static' spreadsheets with formatting (fonts, colours, styles, cell width, etc.), no significant loss of information would occur if the spreadsheets were migrated, to e.g. TIFF or PDF/A. In short, 'simple/static' spreadsheets can be migrated to non-spreadsheet specific file formats or formats that are not meant to preserve dynamic behaviour.

We defined 'complex/dynamic' spreadsheets as spreadsheets that contain formulae, notes, macros, dates, links to external data sources or other functions or dynamic behaviour. Migrating to non-spreadsheet formats could cause severe information loss.

There can be other ways to categorize spreadsheets in a collection, for example we considered treating every file format (version) as a separate sub-type and differentiating between the number of worksheets. If we were to use (versions of) file formats as sub-types, we would be limiting ourselves to comparing technical features of file formats and how well they can be migrated to each other, rather than the significant properties of the more generic high-level object type spreadsheets. The number of worksheets and their interrelatedness can be mapped as (e.g. structural or behavioural) spreadsheet properties and does not really affect the function of the spreadsheet for users. As a result, we decided not to pursue these particular sub-type distinctions.

Spreadsheet Complexity Analyser

The first results of applying the characterisation tools on spreadsheets indicated that they might not extract important spreadsheet-specific properties that we encountered in the spreadsheet specifications. We therefore looked for spreadsheet-specific characterisation tools but did not find suitable (open source) tools.

A second reason to start thinking about developing our own tool was the introduction of the sub-types. In addition to extracting properties, we wanted to have a tool that could be used to help distinguish between our sub-types.

We therefore developed a Spreadsheet Complexity Analyser (SCA), voted on which properties this tool should be able to extract, and decided when a spreadsheet would be deemed simple/static or complex/dynamic. The resulting open-source tool currently extracts values of Microsoft Excel (*.xls and *.xlsx) spreadsheet properties and calculates a spreadsheet complexity assessment based on threshold values.

The threshold values that are used to distinguish between the sub-types are best effort values, discussed by the AIG members. We are aware of the fact that they may not suit everyone's purposes. Different organisations or projects may have different preservation policies or quality control requirements. The SCA therefore explains that the sub-type assessment is tentative, and comes with a configuration file in

which users can define their own thresholds. But even if the SCA is not used to distinguish between these sub-types, it can still be used for characterisation.

The next table shows which properties the SCA currently extracts from spreadsheets and how it distinguishes between sub-types by default.

Property	Simple/Static	Complex/Dynamic
File name	Not used	Not used
File size (in kB)	Not used	Not used
Creation datetime (if available)	Not used	Not used
Last accessed datetime (if available)	Not used	Not used
Last modified datetime (if available)	Not used	Not used
Number of worksheets	<=1	> 1
Number of fonts used	<=1	>1
Number of defined names	<=1	>1
Number of cell styles used	<=1	>1
Number of formulas	0	>0
Number of hyperlinks	0	>0
Number of comments	0	>0
Number of (VBA) macros	0	>0
Number of shapes	0	>0
Number of dates	0	>0
Number of cells used	<=1000	>1000
Number of external links	0	>0
Does the workbook have a revised history	Not used	Not used

3.1.2. Analyse structure

The InSPECT methodology suggests that the evaluator analyses the object to obtain its technical properties. As AIG, we decided to use both analysis methods referred to in the InSPECT methodology: analysing a representative sample of spreadsheets with characterisation tools and review technical spreadsheet specifications.

The result of both options was a list of properties. The next paragraphs and the Results section contain more information on how we established the list and how it was used in subsequent phases in the InSPECT methodology.

3.1.3. Identify purpose of technical properties

The purpose of this activity is to determine the role that the property performs within spreadsheets. We initially used the InSPECT property categories:

1. "Content: Information contained within the Information Object. For example, text, still and moving images, audio, and other intellectual productions. Examples: duration, character count.
2. Context: Any information that describes the environment in which the Content was created or that affects its intended meaning. Examples: Creator name, date of creation.
3. Rendering: Any information that contributes to the re-creation of the performance. For example, font type, colour and size, bit depth.
4. Structure: Information that describes the extrinsic or intrinsic relationship between two or more types of content, as required to reconstruct the performance. E.g., e-mail attachments.
5. Behaviour: Properties that indicate the method/s by which content interacts with other stimuli. For example, hyperlinks, macros."[23]

By assigning every property to one of these categories, an overview was created of the role of the properties. However, with over 400 properties at this point, we noticed two things: (1) even with the categories, the lists of properties per category were too long for practical use and (2) many properties were related to spreadsheet features such as hyperlinks, formulas, table formatting or localisation.

We therefore introduced the concept of property groups and grouped properties together under these groups. This made our work more efficient, as one decision w.r.t a group resulted in a decision for all the properties that fell under that group (it was still possible to make different decisions for individual properties later.) For example: the categories show that the property Table Style belongs to the category Rendering, since it involves the visual look of the spreadsheet. With the property groups, it also fell into the property group Tables.

Property	Category	Property Group
Table Style	Rendering	Tables

With the introduction of the property groups, it became easier to reference and structure the properties and relate them to purpose, behaviours, and functions in the next steps in the InSPECT methodology. Moreover, having these more specific groups turned out to also be useful for the stakeholder analysis. Talking about property groups that more or less reflect spreadsheet features appeals more to the imagination of the stakeholder than the InSPECT categories or the individual properties would.

3.1.4. Determine expected behaviours

To determine expected behaviours, we conducted a joint brainstorm on possible use cases any user could be expected to carry out when working with spreadsheets, or when a consumer wishes to access spreadsheets for reuse from an archive. Examples of use cases 'Understand how data was entered', 'Reproduce charts and (pivot) tables' and 'Investigate accuracy of calculations'. The latter was a use case that resulted from our contact with Felienne Hermans,[24] who developed techniques in detecting errors in spreadsheets.
This list of behaviours was the basis for the next steps of the InSPECT methodology.

[23] A. Wilson, "Significant Properties Report."

[24] "About," Felienne, https://www.felienne.com/about.

3.1.5. Classify behaviours into functions

The various behaviours were classified into a shorter list of spreadsheet functions. These functions also were the result of AIG brainstorms.

3.1.6. Associate structure with each behaviour

In this step of the InSPECT methodology, the list of properties, the property groups and the behaviours and functions were all linked up. We soon referred to the resulting Functions, Structures and Behaviours diagram as our 'spaghetti diagram', because it soon became a complex diagram with many relations.

This linking up of all information was not a one-off process, it was a process with several iterations. While doing this, we also learned that while our property group idea was helpful, the property groups themselves could be improved upon.

3.1.7. Review and finalise

Different from InSPECT we made groups of properties to keep the overview. We re-evaluated our property list and added our opinions about the relevance of properties. 'Relevant' properties were those properties that we as AIG and archive stakeholder found relevant to consider as significant properties. These opinions were useful, because it allowed us to compare our relevance hypothesis to the results of the stakeholder analysis later.

3.2. Stakeholder Analysis

3.2.1. InSPECT framework

The InSPECT methodology suggests to carry out a stakeholder analysis: "The objective of the stakeholder requirements analysis is to identify the stakeholder categories that may have some relationship with the object type/sub-type and determine the set of functions that they require when using it. The set of functions associated with the stakeholder may be subsequently cross-matched with the object type functions and a list of significant properties developed for each context."[25]

3.2.1.1. Identify stakeholders

In order to establish what type of stakeholder is eligible to participate in this research, we have to look at the population. The population is the designated community mentioned in the OAIS model. This community consists of stakeholders that should have no trouble in understanding information that the archive has preserved. Stakeholders should ideally represent entities related to the preservation lifecycle as denoted by the OAIS and may include:

- Archive creators - providers/producers (government, agencies, research institutions etc.)
- Archive curators - memory institutions, archives internationally, nationally, and locally

[25] Significant Properties, "InSPECT Framework Report."

- Users - can be the archive creators, a researcher, or as such in general a recipient of a requested DIP
- Technology providers such as Microsoft and Apple
- Consultants

For this report, a further restriction is made by limiting the population to merely individuals that were employed in the public sector. This is due to the fact that the organisations for which this study is carried out, the National Archives of various countries, preserve information from public institutes. This excludes technology providers from the scope of this report.

3.2.1.2. Select object type for analysis

This report considers spreadsheets created by public authorities and appraised for preservation either directly on an individual basis or indirectly through the total preservation of a records management system. In essence, the latter has the consequence that we need to consider all spreadsheets stored in a records management system and not only those explicitly marked as being of archival value.

3.2.1.3. Remaining steps

The remaining steps that need to be carried out according to the InSPECT framework are the following:

- Determine actual behaviours
- Classify behaviours into set of functions
- Cross-match functions
- Assign acceptable value boundaries

The first step is to determine how a certain type of stakeholder uses the spreadsheet. Concerning the object type email this could for example be viewing the textual content of the message or establishing the email account from which the message originated. With spreadsheets, however, this is an immense task. Stakeholders use spreadsheets for a wide range of activities with no established set of functions that has to be used every time. Furthermore, we felt this would be difficult to accomplish thoroughly during interviews with stakeholders, considering the size of the task. Therefore, these last steps were not performed by us during this research.

3.2.1.4. Review and finalise

After reviewing the InSPECT methodology, we decided to deviate 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, which were less successful, and to come with more extensive recommendations. Furthermore, all three of the National Archives in the AIG had different aims for their case studies. In our opinion, this diversity would enrich the research more by having different views and perspectives.

What we did do as an alternative to the 'Cross-match functions' and 'Review and finalise' steps is to combine the Object Analysis and Stakeholder Analysis results to establish which property groups and properties are seen as significant by the stakeholders. Due to our hypotheses ('relevant' or not) we were also able to determine how well we as archive stakeholders were able to predict these outcomes.

3.2.2. Case studies

3.2.2.1. National Archives of the Netherlands[26]

The Object Analysis yielded an extensive list of 334 properties. Asking stakeholders to select a few out of these would be impractical. It was therefore found imperative to add a type of grouping. Herein lay two options. The first option was provided by the InSPECT Framework, which makes use of categories of behaviour. In total, there are five categories:

1. Content: information content within the spreadsheet. Examples of this are text and images.
2. Context: describes the environment in which the spreadsheet was created and has an influence on its intended meaning. Examples are the initial creator and creation date.
3. Rendering: has an influence on how the spreadsheet looks. Examples are font colour and font size.
4. Structure: describes how two or more types of content are related to each other. Examples of this are auto calculation and cell references.
5. Behaviour: the properties that demonstrate how the content interacts with other stimuli. An example of this are hyperlinks.

However, these five categories are rather broad. For the stakeholder analysis, having a more specific grouping could be more beneficial. Moreover, the categories are quite abstract and do not match the terminology used by the stakeholders themselves. Therefore, the properties were subsumed into 21 groups:

Application Settings	Editing	Macros
Cell Content	External Data	Metadata
Cell Formatting	Formatting	Pivot Tables
Charts	Formulas	Printing
Comments	Graphic Elements	Protection
Data Compression	Hyperlinks	Statistics
Data Tools	Localization	Tables

Using groups also created a better oversight of the function of the property itself.

After subsuming the properties into groups, stakeholders were found to participate. The National Archives of the Netherlands started by setting up three types of roles: maker, user, and manager. This is in line with the InSPECT Framework Report, where there are two requirements set to perform the analysis. The first one concerns the role of the stakeholder, there needs to be a clear understanding as to what the relationship of the stakeholder is with the digital object, in this case, the spreadsheet. The second requirement concurs with this by stating that there need to be multiple stakeholders in each role. The National Archives of the Netherlands questioned 16 stakeholders, of whom seven were employed

[26] For a more extensive report on the Stakeholder Analysis conducted by the National Archives of the Netherlands see: L. Wijsman, "The Significant Properties of Spreadsheets: Stakeholder Analysis," Zenodo, https://doi.org/10.5281/zenodo.3971833.

by various Dutch ministries, whilst the other nine are working for semi-governmental institutes. They range from policy advisors to consultants and specialists. Their knowledge of spreadsheets is diverse because of their diversity in function.

In order to find which properties were deemed significant by the stakeholders, three parts were carried out, which together form a toolbox and methodology that can be applied by archives for future research. The first part consisted mainly of exploratory questions. Examples of these questions were why these stakeholders use spreadsheets and how they qualify their own level of knowledge. In addition, the stakeholders were requested to come up with five properties that seemed important to them when it came to preserving spreadsheets. Furthermore, the stakeholders were asked to submit a representative spreadsheet. This spreadsheet could then be assessed at face-value using the Spreadsheet Complexity Analyser.

The second part was more in-depth. The aforementioned 21 groups were presented in the form of a catalogue.[27] Participants in the study were asked to choose the five groups that they deemed to be most significant. On the basis of these two parts, a follow-up interview was conducted with five participants. These interviews focused more on the background and preservation intent of the stakeholder. Based on the gathered information, qualitative and statistical analyses were carried out.

3.2.2.2. National Archives of Estonia

The National Archives of Estonia interviewed two producers in January 2020: the National Archives of Estonia (NAE) itself and the Estonian Research Council (ETAg). The ca. 45 min interviews with the document manager (archivist) of the organization and IT support (together, not 2 interviews separately) were conducted in the offices of the producers.

The aim of the visits was to look at different spreadsheets used in the operation of the organization today and during the last decade, to get an understanding of the life cycle of these files, and detect any outstanding properties that we might have overlooked in our work in the AIG so far. Quantities such as number of spreadsheets, worksheets, rows; file sizes, etc. were not so much focussed on. Only files that are part of records of archival value were of interest - only the files that will be part of the collection of NAE one day.

A questionnaire was sent to the interviewees beforehand to let them see what the talk would be about. During the interview the questionnaire was not followed strictly at all, the conversation was let to flow to let the stakeholder express what is important in their role, several questions from the questionnaire were not asked at all. Notes were taken on paper and the interviewees were not asked to submit any information in written form.

3.2.2.3. Danish National Archives

We examined spreadsheets in 2019 as a pilot test of a new concept model that we have developed in-house. The purpose of the concept model is to create a framework for developing preservation plans for

[27] L. Wijsman, "Catalogue Significant Properties of Spreadsheets," Zenodo, https://doi.org/10.5281/zenodo.3902080.

content information objects and one of its methods is a "migration assessment", where we apply an adapted version of InSPECT.

The migration assessment analyses information properties of formats and juxtaposes these to the use case of stakeholders such as data producers and archival users. In essence, we adapted InSPECT. The team examining spreadsheets was composed of two academic staff spending close to three full time months on the entire pilot test whereof approx. one month was spent on the migration assessment.

For the stakeholder interviews we sent the questionnaires in advance of the interview and the actual interview was, if feasible, conducted on location of the stakeholder's workplace. If distances were too great to travel, we did an online video meeting instead. At the interviews we would represent with the two staff working on the migration assessment and the stakeholders would represent between 1-3 staff for the interview, which would usually last 1-2 hours.

We interviewed:

- The Head of Finance Department in one of our municipalities
- Two young professionals at the national bank of Denmark (as a curiosity, one of them had participated in the Danish championship for spreadsheets. We didn't even know such a thing existed!)
- The archive NEA which archive data from a network of municipalities
- The archive KOMDA which archive data from a network of municipalities

We also contacted other data producers requesting an interview but received no reply.

All of the interviews went well, and we received important knowledge concerning the use cases and needs for preservation from the data producers and feedback on the issues with our current preservation specification from the municipal network archives.

Our experiences from the stakeholder interviews were that it can be extremely time and competency consuming to analyse every single property and behaviour for a complex content information type such as spreadsheets. In fact, for these kinds of analyses it can be counterproductive to conducting the interview if we do not try to stray away from the InSPECT approach and instead focus on facilitating a meaningful conversation with people and from this conversation try to deduce the behaviours necessary to preserve for future reuse of the data. The questions you instead can ask people are what they deem important to be able to do with the data and what data and associated functionality do they find important to preserve, if they were to reuse it in the future.

4. Results

4.1. Object Analysis

4.1.1. Select object type for analysis

As AIG, we combined spreadsheets from various sources to create a representative sample. We made use of publicly available spreadsheet samples from:

- EUSES[28]
- Enron[29] (spreadsheets only):
- Govdocs[30] (spreadsheets only),
- OPF Format Corpus[31]
- Apache OpenOffice Spreadsheet Test Documents[32]

We also added spreadsheets from our national contexts:

- The National Archives of Estonia shared some publicly available spreadsheets with the AIG members.
- The National Library of the Netherlands offered to analyse 180,000 spreadsheets from their repository using the Spreadsheet Complexity Analyser. The results were shared with the OPF AIG members in private.
- The Danish National Archives ran the SCA against about 16,000 Microsoft Excel spreadsheets (both binary formats and OOXML) to investigate the possible information loss when converting Excel spreadsheets to ODS. Due to confidentiality issues, these spreadsheets could only be used within the Danish National Archives. The test showed that the conversion from XLS and XLSX to ODS and back to XLS and XLSX resulted in minimal data loss. Yet, data loss for significant structures such as cell typographies, fonts and hyperlinks were encountered. The tool could not analyze XLSX Strict, only the transitional equivalent.

4.1.1.1. Technical specifications

We collected technical specifications of specific spreadsheet formats. Many 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. Often several versions exist for each format.

VisiCalc was the market leader until the middle of the 1980s when Lotus 1-2-3 and Borland Quattro took over, and around the middle of the 1990s, Microsoft Excel dominated the market. Excel has since

---

[28] "Modified EUSES Corpus," Spreadsheets, http://spreadsheets.ist.tugraz.at/index.php/corpora-for-benchmarking/euses/.

[29] F. Hermans, "Enron Spreadsheets and Emails," Figshare dataset, https://figshare.com/articles/Enron_Spreadsheets_and_Emails/1221767.

[30] Digital Corpora, http://downloads.digitalcorpora.org/corpora/files/govdocs1/zipfiles/.

[31] "Format corpus," GitHub, https://github.com/openpreserve/format-corpus.

[32] "Spreadsheet Project – Filter Test Documents," Apache OpenOffice, https://www.openoffice.org/sc/testdocs/.

achieved an almost monopoly. In the 2010s we have seen cloud services supplying their own spreadsheet apps, most notably Google. Currently, the most commonly used formats are OpenDocument Spreadsheet and market leader Microsoft with the Office Open Spreadsheet ML (est. above 90%). The older Microsoft XLS (binary) file formats are still in use too.

4.1.1.2. Characterisation tools

As explained in the Methodology chapter, the AIG used FITS, fido, Siegfried, Lingfo (XLRD), Dependency Discovery Tool, Officeparser.py, Ssconvert, Python oletools, Apache POIfs, Apache Tika and (counted as one) some Python libraries to characterise spreadsheets. Examples of properties extracted by these tools (from different spreadsheets) are listed in the next table, together with our assessment of their categories:

Property	Example	Category	Extracted by (tool)
PUID	Fmt/189	Structure	FITS
Size	32658 byte	Context	FITS
Has_hyperlinks	Yes	Structure	FITS
Heading Pairs	Nimega vahemikud, 17	Content	ExifTool
Code Page	Windows Baltic	Rendering	ExifTool
Company	Ernst & Young	Context	ExifTool
CharacterSet	ISO-8859-1	Rendering	New Zealand Metadata Extraction Tool
Creator	Einike	Context	Apache Tika

Different tools have different names for properties, such as “Application-Name” (Apache Tika), “Application” (ExifTool) or “Creating_application_name” (FITS). This is one of the reasons that our initial list of 400 extracted properties was halved after de-duplication and clean-up.

Accompanying the selection of the object type was our division of spreadsheets into two sub-types: simple/static and complex/dynamic. However, after the KB’s analysis of 180.000 spreadsheets with the prototype of the SCA, we saw that almost all (99%) spreadsheets were assigned the label complex/dynamic. This was probably due to our definition of complex/dynamic and our resulting default threshold values of the tool. For example, when a spreadsheet makes use of more than one font, which is often the case, it is instantly labelled as complex/dynamic. This led to the addition of the configuration file to the SCA. It allows users to override the default thresholds.

For future research into dividing spreadsheets into these subtypes we recommend a new definition of when something is considered to be simple/static or complex/dynamic. Also, more research needs to be done to establish SCA thresholds, which might lead to a better assessment of the subtypes. Furthermore, there needs to be a clear assessment of the viewpoint from which this definition is made. For example, is the goal to make a decision between conversion of a spreadsheet format to a TIFF format or to a spreadsheet-specific format.

4.1.2. Analyse structure

As AIG, we decided to use both analysis methods referred to in the InSPECT methodology:

1. Use characterization tools to analyse and extract information on the technical composition of the object for storage as Representation Information.
2. Review the technical specification or standards associated with the object type and identify the technical information that is used to construct the Data Object.

The result of both options was a list of properties. The spreadsheet in which we stored the de-duplicated and cleaned list has a blue title or header row. We therefore soon referred to this spreadsheet as our 'blue sheet'.[33] This list contains 198 properties.

4.1.3. Identify purpose of technical properties

The initial list of property groups was based on the well-known significant property categories content, context, rendering, structure, and behaviour. As we started working towards connecting the properties to purpose, behaviour and function, group names that reflected those characteristics were introduced too: e.g., security for any spreadsheet security-related properties and character formatting for character and cell formatting properties.[34]

When we returned to our list of properties in one of our later iterations, 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 we must credit Frederik Holmelund Kjærskov of the Danish National Archives for proposing to use the 'industry standard' property groups from Apple.[35] We therefore introduced these property groups in the stakeholder analysis work, and especially Lotte Wijsman used this terminology when she conducted her stakeholder analysis in the Netherlands. As we didn't want to delay our internal in-progress object analysis work, we decided not to change the object analysis property groups, but a mapping from the 'old' to the 'new' property groups was available.

One drawback of this 'fork' in our work was that while the Object Analysis continued to work towards the aforementioned 198 properties in the blue sheet, the Dutch Stakeholder Analysis work started from an earlier version of the blue sheet and ended up with 334 properties. Many of the additional properties in this list are properties specific to either Microsoft Excel or Open Document Spreadsheets. The property Accounting format is e.g. a Microsoft-specific property that enables formatting currency information in an accounting-friendly manner. This representation can be reproduced in Open Document Spreadsheets, but it is not part of the specification. Also, the property groups used in the Object Analysis and the Dutch Stakeholder Analysis differed slightly. We therefore established and maintained mapping tables between the versions.

---

[33] See Appendix B.

[34] See Appendix B.

[35] "Document compatibility with Microsoft Office," Apple, https://www.apple.com/mac/numbers/compatibility/.

4.1.4. Determine expected behaviours

As explained in the Methodology section, we conducted a joint brainstorm on spreadsheet use cases, or as InSPECT put it, on: "the different types of activities that a user – any type of user– may wish to perform. The list of activities should be recorded as a set of expected behaviours."

We soon realised that we would never establish an exhaustive list of all possible behaviours and chose to use those behaviours that we found most important from our perspective as archives preserving spreadsheets. The behaviours, and the functions connected to them (see next section) resulted in a spreadsheet with a green title bar, hence a 'green sheet' with behaviours and functions.[36]

4.1.5. Classify behaviours into functions

Similar to how we established the list of behaviours, we brainstormed on functions connected to the behaviours from the previous section. In InSPECT terminology: we "classif[ied] the set of behaviours identified in the previous stage into a set of functions. The functions may be used as a basis for tailoring future manifestations of the Information Object to the needs of the stakeholder."[37]

After revisiting and discussing the expected behaviours and functions several times, we decided on using the following table.

Expected behaviours	Functions
Inspect data dependencies to other sources	Determine data dependencies
Determine relations between worksheets	Determine data dependencies
View changes tracked (hidden history of creation)	Determine privacy issue
View author	Establish context
View creation date	Establish context
Understand the purpose	Establish context
See comments/notes (of a cell)	Establish context
Determine spreadsheet life cycle	Establish usage
Identify the spreadsheet users	Establish usage
Understand the spreadsheet use	Establish usage
Identify the spreadsheet version	Establish version
Inspect the significance of custom formatting	Inspect data rendering
Inspect date/weight/monetary/... formats	Inspect data rendering
Investigate accuracy of calculations	Investigate provision of data
Determine the creating application	Investigate provision of data
Inspect data calculations	Investigate provision of data

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

[37] Significant Properties, "InSPECT Framework Report."

Understand how data was entered	Investigate provision of data
Inspect macros in spreadsheet	Investigate provision of data
View data in cells	Reuse data
View worksheets	Reuse data
Export data to other application	Reuse data
Select subset of data	Reuse data
Interact with interactive content	Reuse graphical objects
Reproduce charts and (pivot) tables	Reuse graphical objects

4.1.6. Associate structure with each behaviour

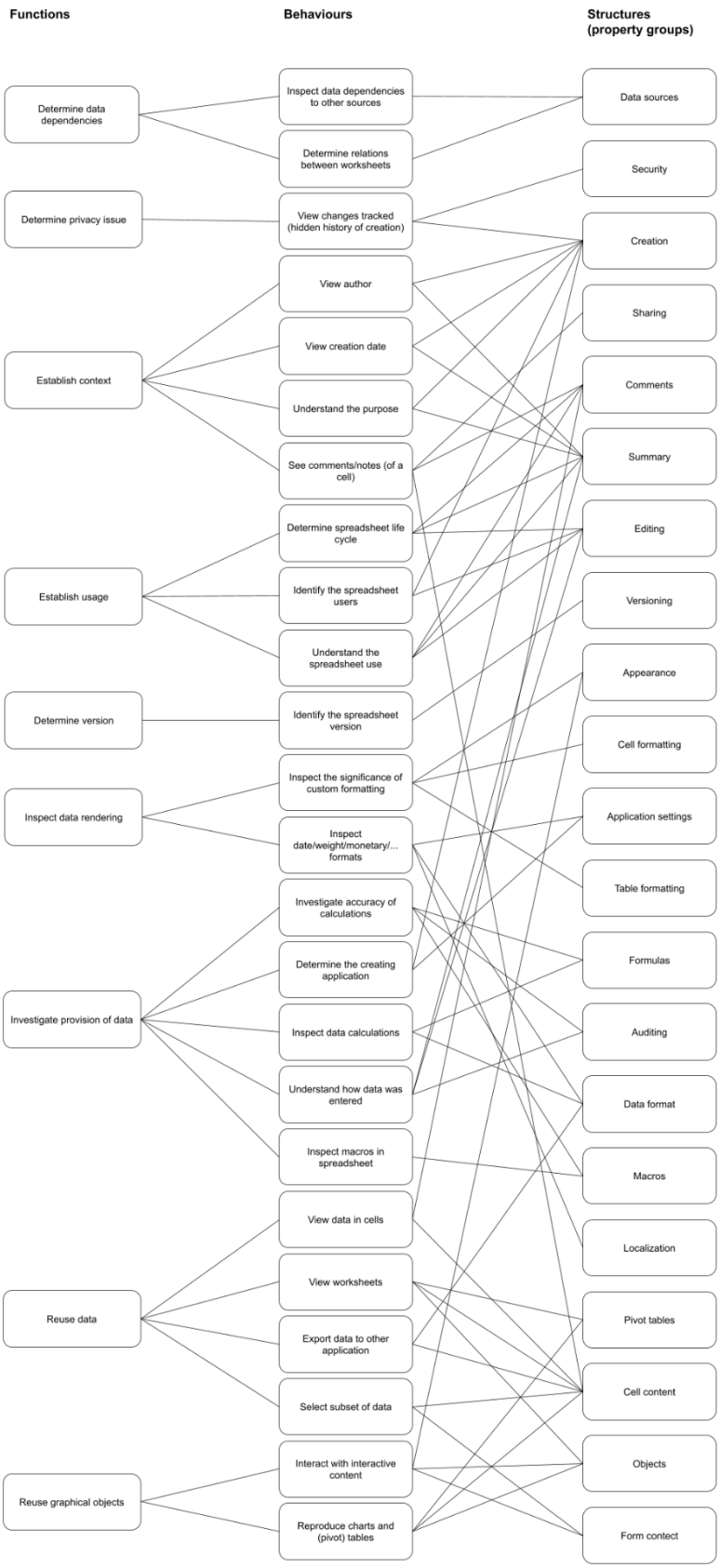
What is missing from the table of the previous section are the properties - or in our case the property groups. That is the main work of this phase of the InSPECT methodology: "link the technical properties that establish the structure of the Data Object with the set of expected behaviours".[38]

This is why we added to our green sheet (with behaviours and functions) columns for associating up to three property groups with the behaviours and functions. The property groups were selectable from a drop-down list, populated from the list of property groups from the blue sheet.

As a result, we were now able to create a Function-Behaviour-Structure (or functions, behaviours, property groups) diagram. As mentioned before, this diagram was soon referred to as our spaghetti diagram. The diagram is included below and available online.[39]

[38] Significant Properties, "InSPECT Framework Report."

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


Functions
Behaviours
Structures (property groups)
Determine data dependencies
Determine privacy issue
Establish context
Establish usage
Determine version
Inspect data rendering
Investigate provision of data
Reuse data
Reuse graphical objects
Inspect data dependencies to other sources
Determine relations between worksheets
View changes tracked (hidden history of creation)
View author
View creation date
Understand the purpose
See comments/notes (of a cell)
Determine spreadsheet life cycle
Identify the spreadsheet users
Understand the spreadsheet use
Identify the spreadsheet version
Inspect the significance of custom formatting
Inspect date/weight/monetary/... formats
Investigate accuracy of calculations
Determine the creating application
Inspect data calculations
Understand how data was entered
Inspect macros in spreadsheet
View data in cells
View worksheets
Export data to other application
Select subset of data
Interact with interactive content
Reproduce charts and (pivot) tables
Data sources
Security
Creation
Sharing
Comments
Summary
Editing
Versioning
Appearance
Cell formatting
Application settings
Table formatting
Formulas
Auditing
Data format
Macros
Localization
Pivot tables
Cell content
Objects
Form contect

4.1.7. Review and finalise

Where InSPECT proposes to "review the information gathered in the previous steps and consider if any revisions should be made", we already mentioned that "linking up of all information was not a one-off process, it was a process with several iterations". So instead of having one meeting to review and finalise our results, we revisited and discussed our results in monthly meetings, on our email list and in one or two ad hoc cooperative sessions. But similar to InSPECT, we decided after several iterations that our results were 'good enough'.

In one of the iterations, we introduced a Status column in the blue sheet. This column gave us the opportunity to assign a 'Relevant', 'Not relevant', 'Investigate', 'Not linked to behaviour' and 'Keep or remove?' status to properties. 'Relevant' properties were those properties that we as AIG and archive stakeholder found relevant to consider as significant properties. Not relevant properties were not considered as significant. We could also keep track of any properties that were not (yet) linked to behaviour (as relevant or not relevant). Some properties needed more investigation, e.g. by looking up more information about them in specifications. And there were also properties that we considered redundant. Those required a discussion about keeping or removing

4.2. Stakeholder Analysis

4.2.1. National Archives of the Netherlands

After the exploratory questions, the stakeholders were assigned groups with regard to their gender,[40] level of knowledge, and role. The results of this can be seen in the table below:

Stakeholder	Gender	Knowledge	Role
1	Male	Advanced	Maker/user
2	Male	Advanced	Maker/user
3	Male	Advanced	Maker/user
4	Male	Advanced	Maker/user
5	Male	Average	Maker/user
6	Female	Average	Maker/user
7	Male	Average	Maker/user
8	Female	Basic	Maker/user
9	Male	Advanced	Maker/user
10	Male	Advanced	Maker/user
11	Male	Average	Manager
12	Male	Average	Maker/user
13	Male	Advanced	Maker/user
14	Female	Advanced	Maker/user

[40] The specified genders were selected by the stakeholders themselves and were based on how the stakeholders represented themselves.

15	Male	Advanced	Maker/user
16	Female	Basic	Manager

As shown in this table, the roles of maker and user have been merged. This is due to the fact that the stakeholders often indicated that they deemed the two roles to be intertwined. Therefore, the two groups were fused together as one.

After the preliminary questions were asked, the stakeholders filled out the reply form of the catalogue that was created. This resulted in the following choices being made by the stakeholders:

Property Groups	**Stakeholders**															
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
Protection		x						x		x			x			
Editing						x										x
Cell Content	x			x			x		x	x	x	x		x		x
Cell Formatting			x	x							x	x		x	x	
Data Tools						x		x	x							
Pivot Tables	x	x	x	x	x					x				x	x	
External Data			x	x	x	x			x	x	x		x	x	x	
Formulas	x	x	x			x	x	x	x	x	x	x		x	x	x
Charts		x	x	x	x	x						x			x	
Graphic Elements							x									
Hyperlinks							x									
Macros								x					x			
Metadata	x						x				x	x	x			x
Formatting	x				x			x	x							
Comments																x
Statistics																
Tables		x			x											
Data Compression																
Localization																
Printing																
Application Settings													x			

The five most selected groups were:

- Formulas (chosen 13 times). A formula calculates the value of a cell (or multiple cells). For instance, the formulas AVERAGE and SUM. Some properties in this group are formulas, financial functions, custom calculation, statistical functions, and subtotal.
- External Data (chosen 10 times). This is data that exists outside of the application itself. The external data is retrieved by the application from an external source via queries. This data may change over time, it is often dynamic. Some properties in this group are DDE (Dynamic Data

Exchange) connections, OLE (Object Linking and Embedding) objects, table DDE links, and web queries.

- Cell Content (chosen 9 times). Cell content is (for the purpose of this study) any text you store in a cell. This group has only one property, namely basic text content.
- Pivot Tables (chosen 8 times). A pivot table is a table that summarises the data of a more extensive table into key statistics, such as the mean and sums. Some properties in this group are pivot table, calculated fields, grouping, and layout.
- Charts (chosen 7 times) A chart lets you visually display data in various types of charts, such as bar, column, and pie. Some properties in this group are bar chart, pie chart, chart layout, and legends.

From the five deemed most significant, four are dynamic property groups. In certain cases, reasons for why a group was not selected as significant could be found in the results of the explorative study. For instance, one of the stakeholders that did not select formulas indicated in these questions that their level of knowledge was average and that they did not want to work with formulas since they were not comfortable with more advanced functionalities of spreadsheets. Their submitted representative spreadsheet also showed an absence of formulas. This also showed the importance of having the representative spreadsheets and being able to assess these at face-value using the Spreadsheet Complexity Analyser.

Rather than simply finding which groups were deemed most significant, the research also wanted to explore if there were patterns to be found by combining results. Using STATA,[41] several analyses could be made of which two will be laid out here. Both analyses study the influence of knowledge and gave statistically significant results. The first analysis was a tabulation of the property group pivot tables and stakeholder knowledge.

Knowledge	**Pivot Tables**		
	No	Yes	Total
Basic	2	0	2
	100%	0.0%	100%
Average	4	1	5
	80%	20%	100%
Advanced	2	7	9
	22.2%	77.8%	100%
Total	8	8	16
	50%	50%	100%
	Probability		0.037

This table shows how level of knowledge might influence a stakeholders choice to consider pivot tables to be significant. A probability of 0.037 means that there is a 96.3% chance that a higher level of knowledge indeed leads towards a higher percentage of choosing the property group pivot tables to be significant.

[41] STATA is software for statistics and data science. STATA, https://www.stata.com/.

The second analysis is a tabulation of level of knowledge and the role of the stakeholder.

Role	Knowledge			
	Basic	Average	Advanced	Total
Maker/user	1	4	9	14
	7.1%	28.6%	64.3%	100%
Manager	1	1	0	2
	50%	50%	0.0%	100%
Total	2	5	9	16
	12.5%	31.3%	56.3%	100%
		Probability		0.128

The results show that makers/users assign themselves with a higher level of knowledge. The probability of differences between the two groups is 87.2% However, here a limitation is seen. Having 16 stakeholders is sometimes not enough to have conclusive evidence. Having more stakeholders in every group could lead to results with a lower *p value*, resulting into more conclusive findings.

Concluding, the Stakeholder Analysis conducted by the NANETH had several findings. The first finding was the importance of exploratory questions. Having information on e.g. level of knowledge and role helps determine certain patterns. Furthermore, assessment of the representative spreadsheets with the Spreadsheet Complexity Analyser helped to get an objective view of what the stakeholders encounter in their work. Having this background knowledge prior to the interview helps the researcher come well prepared, thereby allowing them to use the interview to further clarify certain aspects.

It is important to clarify that there is no 'one size fits all' solution. Every stakeholder uses spreadsheets differently and deems other properties to be of significance. Therefore, establishing patterns and best practices is needed. By doing more research on this and having a bigger sample, this can be achieved.

4.2.2. National Archives of Estonia

NAE focussed the interviews on these spreadsheets that were registered to the series that have appraised as being of archival value during the past decade.

Nowadays only Microsoft Excel is used as creating and editing application for spreadsheets but earlier also LibreOffice Calc was often used. Ca. 7-10 people create spreadsheets of archival value in both organizations.

One stakeholder reported that today ca 10 spreadsheets a year are created, not more. Later checking it in the electronic records management system (ERMS) revealed a much bigger number. It may be due to a misunderstanding: the question was about (technical) files, but they interpreted it as documents (records) that may consist of several files of different formats.

Approximately half of the files could not have been created in any other format than a spreadsheet format, among the reasons were dynamic content and template given from the ministry.
Over the decade, a lot of information of archival value has been moved from spreadsheet files to several national registries and will be archived as databases.

An interesting finding was the presence of files created only recently (2016-2019) but in old XLS format. Possible reason for that is that people use ancient templates and nobody has updated the template to produce newer format because the record layout itself has remained intact over time (e.g the same data to be reported every December).

Most important findings:

- To a question about important aspects, a spreadsheet creator replies, "in my spreadsheets, all aspects are important".
- Estimated amount of spreadsheets may differ greatly from the actual number.
- Background colour and text colour definitely bear meaning and therefore are significant properties.
- Interesting habits of employees may reveal (same old templates used over years).
- You may struggle making the stakeholders see what you are asking about and why it matters.
- Usage of spreadsheets is getting smaller due to many bureaucratic procedures having been "moved into" national registries.

4.2.3. Danish National Archives

We performed four separate interviews with two data producers and two archives. 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 the structures necessary to preserve through their eyes and experience.

The interviewed stakeholders pointed to significant properties, which can't be converted without loss to DNA's accepted format for documents, the imaging format TIFF. These properties contribute to the documentation, the correct understanding, and the interpretation of the content in spreadsheets.

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 options for documenting the origin and interpreting the content. This is for instance the case with formulas, references to defined names, the data areas of graphs, number formatting, conditional formatting, and calculated values for pivot charts etc.

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 many hundreds of printed pages is in sharp contrast to the preservation objective of later reuse of data.

Those properties lost by conversion to TIFF are not only significant for the functionality of the spreadsheet but also for the semantic understanding of the content. Therefore, based on our interviews with the four stakeholders, it is our assessment that if a spreadsheet is worthy of preservation, then it is by all means unacceptable to lose properties at the cell level, which can contribute to the understanding of the preserved content.

	Data Producers		Archives		
Structures	National bank	Municipality	NEA	KOMDA	Assess. of significance
Formulas	✓	✓	✓	✓	Significant
Search, sorting and filtering	✓	✓	✓	✓	Significant
Data sources	✓	✓			Significant
Pivot charts	✓	✓			Significant
Context resume				✓	Significant
Sharing	✓				Significant
Embedded objects	✓				Significant
Macros	x	x	x		Insignificant
Page layout					Unknown
Revision					Unknown
Data validation					Unknown
Rendering					Unknown
Typography					Unknown
Application settings					Unknown
Security					Unknown
Localisation					Unknown
Table and cell formatting					Unknown
Number formatting					Unknown

4.2.4. Conclusions

The case studies carried out by the three National Archives tried not only to 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. Concerning the preparation and conducting of the interviews, we found that several things are of importance. As mentioned previously in this report, we felt it would be too difficult to employ the InSPECT methodology for this part due to its abstract nature and the extensive, almost insurmountable, amount of work it would take. 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. One of our findings was that it is vital to have a clear understanding between the interviewer and the stakeholder concerning terminology. Furthermore, the interviewer must create a clear overview for themselves, understanding what they mean with certain terms. This understanding surpasses terminology since it also concerns the background of the stakeholder. As mentioned in the stakeholder analysis conducted by the NANETH, the background of the stakeholder can give a lot of insight into their opinion of what is significant. A stakeholder that never makes use of more advanced features, such as formulas and macros, will not deem these significant. A more knowledgeable user of spreadsheets will however deem these to be of the utmost significance. Therefore, it is important to look at the stakeholders previous work with spreadsheets. An

efficient way to do this is to look at their previous work with spreadsheets by using the Spreadsheet Complexity Analyser. This will help assess the spreadsheets at face-value to see which properties they contain. Learning about the background of the stakeholder would preferably be done before the interview, so the interviewer already has some objective information and can ask where a more substantive answer is needed.

The properties that came forward as being significant during the stakeholder analyses were almost all dynamic in nature. 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 stress here again 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. The results from the interviews immediately confirmed our suspicion we had at the start of this research, that certain current practices are not sustainable and must be revised. Imaging spreadsheets 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 formats such as XLSX and/or ODS could solve the problems that are currently encountered.[42] Moving forward, we recognize that our sample was quite small and more stakeholders need to be interviewed to expand the knowledge-base concerning significant properties. However, we hope that the lessons we have learned and the insights we have resonate with the community and will be employed in the future.

4.3. Combining Object Analysis and Stakeholder Analysis

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. As explained under Identify purpose of technical properties, we had mappings available between the various lists of property groups and properties. An example of this mapping is presented in the next table.

In that table, NANETH SA is short for the Stakeholder Analysis performed in the Netherlands (using a longer list of 334 properties and 'industry standard' property groups). OA is short for the Object Analysis (using a shorter list of 198 properties). The empty cell is an example of a property that is only present in the NANETH SA. For the purpose of filtering, sorting and analysing the data, we copied the NANETH SA property group value to the OA property group in these cases.

NAE SA and DNA SA are short for the Stakeholder Analyses performed in Estonia and Denmark. Significant is used if a particular Stakeholder Analysis mentions that the stakeholders indicated a property group or property as significant, otherwise the significance is Unknown.

[42] The Tagged Image File Format (TIFF) is currently accepted by the DNA, but they are currently working on revising their policy of accepted and preferred formats since this format will not support the migration of dynamic properties. Our research has made clear how significant these are and why they must be preserved.

Properties (NANETH SA)	Property group (OA)	Property group (NANETH SA)	Significance (NANETH SA)	Significance (NAE SA)	Significance (DNA SA)	Properties (OA)
Chart Title	Charts	Charts	Significant	Unknown	Unknown	Charts
Code Page	Localization	Localization	Unknown	Unknown	Unknown	Code page
Codes	Formulas	Formulas	Significant	Unknown	Significant	Codes
Colour	Appearance	Formatting	Unknown	Significant	Unknown	Colour Properties
Column Chart	Charts	Charts	Significant	Unknown	Unknown	
Column Formatting	Formatting	Formatting	Unknown	Unknown	Unknown	Column Formatting Properties

To find 'the' significant properties of spreadsheets, we combined the information from the three Significance columns. The three Stakeholder Analyses were performed in different ways and yielded different results. We wanted to keep our calculations as simple as possible, and without attributing a different value to any Stakeholder Analysis. We therefore chose to mark a property (group) as Significant if any Stakeholder Analysis marked it as significant. In future studies, more elaborate analyses can be performed. For the example of the table above, this results in the next table.

Properties (NANETH SA)	Property group (OA)	Property group (NANETH SA)	Significance (NANETH SA)	Significance (NAE SA)	Significance (DNA SA)	Properties (OA)	Overall significance
Chart Title	Charts	Charts	Significant	Unknown	Unknown	Charts	Significant
Code Page	Localization	Localization	Unknown	Unknown	Unknown	Code page	Unknown
Codes	Formulas	Formulas	Significant	Unknown	Significant	Codes	Significant
Colour	Appearance	Formatting	Unknown	Significant	Unknown	Colour Properties	Significant
Column Chart	Charts	Charts	Significant	Unknown	Unknown		Significant
Column Formatting	Formatting	Formatting	Unknown	Unknown	Unknown	Column Formatting Properties	Unknown

The result of this exercise is a list of all property groups and properties that our stakeholders deemed significant.

According to the Stakeholder Analysis, the significant property groups and properties of spreadsheets are:

Significant NANETH SA property groups	Significant OA property groups	Significant NANETH SA properties	Significant OA properties
Application Settings	Appearance	1904 Date System	1904 date system
Cell Content	Auditing	Annotation	Annotation
Cell Formatting	Cell Content	Area Chart	Auditing tracer arrows
Charts	Charts	Auditing Tracer	Basic Text Content
Data Tools	Context	Arrows	Calculated fields
Editing	Data Format	Bar Chart	Calculated items
External Data	Data Sources	Basic Text Content	Cell references
Formatting	Data Tools	Box and Whisker	Change tracking
Formulas	External Data	Chart	Change Tracking
Graphic Elements	Form Content	Bubble Chart	Metadata
Metadata	Formulas	Calculated Fields	Chart sheet
Pivot Tables	Objects	Calculated Items	Charts
	Pivot Tables	Category Axis Title	Codes
	Range	Category/Series Labels	Colour Properties
	Sharing	Cell References	Connector Properties
		Change Tracking	Consolidation
		Change Tracking Metadata	Custom calculations
		Chart Data Source	Customized error values and empty cell values
		Chart Layouts	
		Chart Sheets	Data Pilot Tables
		Chart Styles	Data Styles
		Chart Title	Date format
		Codes	DDE Connections
		Color	Embedded objects
		Column Chart	External data ranges
		Combo Chart	External links
		Connector	Fill Properties
		Consolidation	Filters
		Cube Functions	Format
		Custom Calculations	Formulas
		Customized Error Values and Empty Cell Values	Grouped items in fields
			Labels in formulas
		Data Labels	Measure Properties
		Data Pilot Tables	Number formats
		Data Styles	Page fields in rows or columns
		Data Tables	
		Database Functions	pivot tables
		Date and Time Functions	PivotTable reports
			Relationships
		Date Format	Share document
		DDE Connections	Shared Workbook information
		Doughnut Chart	
		Drop Lines	Sparklines

	Embedded Objects	Subtotals
	Engineering Functions	Table DDE Links
	Error Bars	Web queries
	External Data Ranges	
	External Links	
	Fill	
	Filter	
	Financial Functions	
	Format	
	Formulas	
	Funnel Chart	
	Grouped Items in Fields	
	Grouping	
	Hi-Low Lines	
	Histogram Chart	
	IMBI PivotTables	
	Information Functions	
	Labels in Formulas	
	Layout	
	Leader Lines on Data Labels	
	Legends	
	Line Chart	
	Logical Functions	
	Lookup and Reference Functions	
	Map Chart	
	Math and Trigonometry Functions	
	Measure	
	Names	
	Number Format	
	OLAP Formulas	
	OLAP Pivots	
	OLE Objects	
	Page Fields in Rows or Columns	
	Pareto Chart	
	Pie Chart	
	Pivot Table Reports	
	Pivot Tables	
	Query Tables	
	Radar Chart	
	Regular Expressions (RegEx)	

	Relationships
	Series Axis Title
	Series Data Source
	Series Order
	Shapes on Charts
	Share Document
	Shared Workbook Information
	Show Data Table
	Show Legend Keys in Data Table
	Show Series Major Gridline
	Show Series Minor Gridline
	Sort
	Spark Lines
	Statistical Functions
	Stock Chart
	Subtotal
	Sunburst Chart
	Surface Chart
	Table DDE Links
	Text Functions
	Treemap Chart
	Trendlines
	Value Axis Title
	Waterfall Chart
	Web Queries
	XY (Scatter) Chart

We also compared the relevance - 'significance' - hypotheses we established as archive stakeholders to the overall significance of the Stakeholder Analyses. Our hypothesis is that we as archive stakeholders are able to determine the significant properties of spreadsheets without consulting other stakeholders.

The calculation of the extent to which our relevance predictions were correct is simple: compare properties we labelled as Relevant to all properties with an overall significance label of Significant. Only if a property is marked as Relevant and as Significant, our hypothesis was confirmed. The following table illustrates the result, using the same examples as before.

Properties (NANETH SA)	Property group (OA)	Property group (NANETH SA)	Properties (OA)	Hypothesis	Overall significance	Hypothesis confirmed/ rejected
Chart Title	Charts	Charts	Charts	Relevant	Significant	Confirmed
Code Page	Localization	Localization	Code page	Relevant	Unknown	Rejected
Codes	Formulas	Formulas	Codes	Relevant	Significant	Confirmed
Colour	Appearance	Formatting	Colour Properties	Relevant	Significant	Confirmed
Column Chart	Charts	Charts			Significant	Rejected
Column Formatting	Formatting	Formatting	Column Formatting Properties	Relevant	Unknown	Rejected

The first result of this exercise is that we can now create a list of confirmed significant property groups and properties. I.e., a list of those that are labelled both Relevant and Significant.

We consider the previous longer lists and the shorter 'confirmed' lists the long list and short list of significant property groups and properties, as established in our investigation by interviewing our stakeholders. There were too many uncertainties in our results to claim that we had found 'the' significant properties of spreadsheets. What we had found were a short list of properties that archive stakeholders and other stakeholders agreed upon and a long list of properties to consider as additional significant properties.

At the property group level, we learned that the Significant and Confirmed NANETH SA property groups were identical. And that mappings showed that all Significant OA property groups matched that list. The property groups Auditing, Data Tools and External Data were not part of the Confirmed significant OA property groups, but that is mostly an artefact of our mappings. In short, we felt that we did find 'the' significant property groups of spreadsheets: Application Settings, Cell Content, Cell Formatting, Charts, Data Tools, Editing, External Data, Formatting, Formulas, Graphic Elements, Metadata and Pivot Tables.

Confirmed significant NANETH SA property groups	Confirmed significant OA property groups	Confirmed significant NANETH SA properties	Confirmed significant OA properties
Application Settings	Appearance	1904 Date System	1904 date system
Cell Content	Cell Content	Annotation	Annotation
Cell Formatting	Charts	Basic Text Content	Basic Text Content
Charts	Context	Calculated Fields	Calculated fields
Data Tools	Data Format	Calculated Items	Calculated items
Editing	Data Sources	Cell References	Cell references
External Data	Form Content	Change Tracking	Change tracking
Formatting	Formulas	Change Tracking	Change Tracking
Formulas	Objects	Metadata	Metadata

Graphic Elements	Pivot Tables	Chart Data Source	Chart sheet
Metadata	Range	Chart Layouts	Charts
Pivot Tables	Sharing	Chart Sheets	Codes
		Chart Styles	Colour Properties
		Chart Title	Connector Properties
		Codes	Consolidation
		Colour	Custom calculations
		Connector	Data Pilot Tables
		Consolidation	Data Styles
		Custom Calculations	Date format
		Data Pilot Tables	DDE Connections
		Data Styles	Embedded objects
		Date Format	External data ranges
		DDE Connections	External links
		Embedded Objects	Filters
		External Data Ranges	Format
		External Links	Formulas
		Filter	Grouped items in
		Format	fields
		Formulas	Labels in formulas
		Grouped Items in	Number formats
		Fields	pivot tables
		Labels in Formulas	PivotTable reports
		Number Format	Relationships
		Pivot Table Reports	Shared Workbook
		Pivot Tables	information
		Relationships	Subtotals
		Shared Workbook	Table DDE Links
		Information	Web queries
		Subtotal	
		Table DDE Links	
		Web Queries	

Another result of this exercise is that, at the individual property level, only 32% of the hypotheses are confirmed. I.e., in 32% of the cases, a property that we as archive stakeholders labelled as Relevant also received the label (of a) Confirmed (hypothesis). If we look at the same analysis from the perspective of the Stakeholder Analysis and compare confirmed hypotheses to properties with an Overall significance label of Significant, the calculation returns 36%. I.e., in 36% of the cases, a property with an Overall significance label of Significant also received the label (of a) Confirmed (hypothesis).

If we perform the same analysis at the level of property groups and label all properties of a property group Relevant or Significant if any property of that group had that label, the percentages are higher. The following table uses the same example as before, but now at this property group level. The differences with the previous table have been coloured.

Properties (NANETH SA)	Property group (OA)	Property group (NANETH SA)	Properties (OA)	Hypothesis	Overall significance	Hypothesis confirmed/ rejected
Chart Title	Charts	Charts	Charts	Relevant	Significant	Confirmed
Code Page	Localization	Localization	Code page	Relevant	Unknown	Rejected
Codes	Formulas	Formulas	Codes	Relevant	Significant	Confirmed
Colour	Appearance	Formatting	Colour Properties	Relevant	Significant	Confirmed
Column Chart	Charts	Charts		Relevant	Significant	Confirmed
Column Formatting	Formatting	Formatting	Column Formatting Properties	Relevant	Unknown	Rejected

As you can gather from this table, the analysis at the property group level results in (many) more properties with a Relevant label and/or a Significant label. The resulting percentages are now 49% (properties with a Relevant and a Confirmed label) and 94% (properties with a Significant and a Confirmed label).

The result of this combination of our results of the Object Analysis and Stakeholder Analysis is, that we think our work demonstrates that performing a(ny) stakeholder analysis is important. As archive stakeholders, we were only able to predict one third to half of the significant properties of the Stakeholder Analysis. Also, where we were unable to claim that we had found the significant properties of spreadsheets, we did find the significant property groups of spreadsheets and short and long lists of properties to consider as significant properties in future investigations.

A spreadsheet with all analysis details is available in a separate file 'Combined (relevant and significant properties)' in .xlsx and .ods format at https://doi.org/10.5281/zenodo.5387099.

5. Conclusion

5.1. General conclusions

After six years of monthly, ad hoc and face to face meetings, individual and group work, seeing group members come and go, stakeholder interviews, update papers and presentations and winning awards, we have brought our investigation of significant properties of spreadsheets to a close.

While it was sometimes difficult to maintain momentum in the work – we are an interest group of volunteers, not a dedicated project team with project funding – we enjoyed cooperating across national and institutional boarders. We learned a lot about spreadsheets and about the group members' different preservation approaches to dealing with them. And had some fun as well.

Gaining knowledge of spreadsheet properties is really helpful when choosing a preservation strategy to preserve spreadsheets. It doesn't matter if this is file format migration, emulation or choosing preferred formats. One example is that the Danish National Archives used the gained knowledge of spreadsheet properties in the decision to revise their accepted formats and adopt a spreadsheet-specific format, which probably will be the Open Document Spreadsheet format.

Performing the Object Analysis stage of the InSPECT methodology resulted in valuable insights into spreadsheets in general and their specifications and (technical) properties in particular. As a side effect, we developed the Spreadsheet Complexity Analyser.

Although we deviated from the InSPECT methodology's workflow for a Stakeholder Analysis – we all simplified it, adapting it to our needs and context of particular stakeholders - we added three reusable ways to interview stakeholders and elicit their opinions on significant properties. What we learned here is that the level of property groups is invaluable for the Stakeholder Analysis. Stakeholders find it difficult enough to talk about significant properties in general, let alone about the significance of hundreds of spreadsheet properties.

Also, where we were unable to claim that we had found the significant properties of spreadsheets, we did find the significant property groups of spreadsheets and short and long lists of properties to consider as significant properties in future investigations.

The most important insight of having performed the Object Analysis and the Stakeholder Analysis is however, that we think our work demonstrates that performing a(ny) stakeholder analysis is important. As archive stakeholders, we were only able to predict one third to half of the significant properties of the Stakeholder Analysis. This may seem obvious, but to the best of our knowledge, it wasn't demonstrated in a larger piece of work of a project or interest group before.

This investigation set out to add a testing report to the InSPECT testing report lore. We drifted away from this goal. The focus was on performing the Object Analysis and Stakeholders Analysis and on what we could learn from them and from the combination of the two. But our work provided the lists, tools and insights DNA required to revise their accepted format policy and adopt a spreadsheet-specific format in their Preservation Policy. Others can also use our work to their advantage.

5.2. Object Analysis conclusions

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 characterisation 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 characterisation tools were deemed to be insufficient. Hence, our research led us to the creation of the Spreadsheet Complexity Analyser, which formed an addition to the other tools by extracting information about cells, sheets, formulas, named objects, macros, etc.

After using the characterisation tools, in the second step of the Object Analysis our research focussed 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 focussed 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.

5.3. Stakeholder Analysis conclusions

Our stakeholder analyses resulted in several findings. By not following the InSPECT methodology due to reasons explained previously, our analyses were all structured differently. This led us to also search for best-practices, since we had found no other stakeholder analysis work concerning significant properties. When we brought our methodologies and results together for this research, we found that several findings had been important to us all. One of these findings concerns background. By establishing a clear overview of the background of the stakeholder before the interview, the interviewer can build a more objective view. Questions such as level of knowledge and role help the interviewer to understand the stakeholder point of view. As seen, people with an advanced level of knowledge are more inclined to use certain properties and also deeming these significant. A tool that can be used to establish this objective view is the Spreadsheet Complexity Analyser. By analysing the (type of) spreadsheets that are often used by the stakeholder, the interviewer can assess the spreadsheet at face-value.

Concerning which properties were deemed significant by the stakeholders, we found that often the dynamic properties are selected, such as formulas and pivot tables. However, this is highly dependent on the stakeholder. Therefore, it is useful to be able to establish patterns and expectations (a higher level of knowledge is more inclined to deem the property group pivot tables more significant) so the interviewer has a footing when entering the in-depth interview with the stakeholder.

Although we are not claiming that the calculations on our combined Object Analysis and Stakeholder Analysis have a high scientific value, we can draw some tentative conclusions. The most important conclusion is that our results demonstrate that performing a Stakeholder Analysis is important. Archive stakeholders need to cooperate with other stakeholders to determine the significant properties of spreadsheets, as 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. At the individual property level, we tend to underestimate what is significant, whereas at the property group level, we overestimate what is significant. I.e., if we don't cooperate with other stakeholders, there is a much higher chance that information loss will occur at some point in time, when we perform some preservation action on spreadsheets.

6. Recommendations

Investigating the significant properties of spreadsheets, or any other type of information, is not easy. There are many variables. And not enough best practices. But, are we not doing this because it is difficult, or is it difficult because we are not doing this? We decided to take the latter view, do this and make it less difficult. We strongly recommend that more knowledge and experiences w.r.t. significant properties are formed and shared. Because if you look closely at your digital preservation strategy, it boils down to finding the best way to preserve the significant properties of information.

An important recommendation to add is that, whatever preservation strategy you adopted, you should include stakeholders in your decision-making process. Even if you don't have the means to preserve all properties that your stakeholders deem significant, you should be aware of which properties those are. Our work demonstrated that we were only able to predict one third to half of the significant properties from our stakeholder analysis.

What we would also like to recommend is that more research is done in the area of the stakeholder analysis itself. Finding the significant properties of spreadsheets is no easy feat. The magnitude of diversity that is present in spreadsheet properties and spreadsheet usage in different contexts makes no investigation the same. By conducting the stakeholder analyses in different ways, we have hopefully provided readers with lists, tools and insights to prepare and conduct their own analyses. Since our results contains uncertainties and our stakeholder sample was relatively small, we strongly encourage people to perform their own analyses. With more research done, we will get closer to community best-practices and a common ground concerning the stakeholder analysis stage of significant properties investigations. This will also enable us to compare our work to others and learn from this.

Follow-up research also needs to be done concerning the selection of preservation strategies and finding suitable formats for preserving spreadsheets and other types of information. We feel that often, the selection of a strategy is steered by what is (subjectively) possible, not what is (objectively) required. Why exactly are certain file formats acceptable or preferred when it is not possible (e.g. because of policy or obsolete formats) to retain information in its original file format? Or: which formats preserve the significant properties best? When you embraced emulation, do you emulate everything or are you also using a migration strategy for some types of information? Again: which strategy is best suited to preserve the significant properties best?

References

Apache OpenOffice. "Spreadsheet Project – Filter Test Documents." https://www.openoffice.org/sc/testdocs/.

Apple. "Document compatibility with Microsoft Office." https://www.apple.com/mac/numbers/compatibility/.

CASPAR Preserves. http://casparpreserves.digitalpreserve.info/.

Dappert, A. & 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).

Digital Corpora. https://downloads.digitalcorpora.org/corpora/files/govdocs1/zipfiles/.

Felienne. "About." https://www.felienne.com/about.

GitHub. "Format corpus." https://github.com/openpreserve/format-corpus.

Hermans, F. "Enron Spreadsheets and Emails." Figshare dataset. https://figshare.com/articles/Enron_Spreadsheets_and_Emails/1221767.

Jones, M. "The Cedars Project." *Library and Information Research* 26, no. 84 (2002): 11-16. http://dx.doi.org/10.29173/lirg136.

Knijff, J. Van der. "PDF/A as a preferred, sustainable format for spreadsheets?" OPF Blog, December 9th, 2016. https://openpreservation.org/blogs/pdfa-as-a-preferred-sustainable-format-for-spreadsheets/.

Lucker, P., C. Sijtsma & R. Van Veenendaal. "Significant Significant Properties – Award Winner: Popular Poster." June 20th, 2019. https://osf.io/rtjw3/.

Microsoft. "Differences between the OpenDocument Spreadsheet (.ods) format and the Excel for Windows (.xlsx) format." https://support.microsoft.com/en-us/office/differences-between-the-opendocument-spreadsheet-ods-format-and-the-excel-for-windows-xlsx-format-3db958c8-e0ac-49a5-9965-2c2f8afbd960?ui=en-us&rs=en-us&ad=us.

PLANETS. https://planets-project.eu/.

Rothenberg, J. & T. Bikson. "Carrying Authentic, Understandable and Usable Digital Records Through Time." Santa Monica, CA: RAND Corporation, 1999. https://www.rand.org/pubs/rand_europe/RE99-016.html.

Shala, L. & A. Shala. "File Formats – Characterization and Validation." *IFAC-PapersOnLine* 49, no. 29 (2016): 253-258.

Significant Properties (archived version). "InSPECT Framework Report." https://web.archive.org/web/20160520083956/http://www.significantproperties.org.uk/inspect-framework.html.

Significant Properties (archived version). "Significant properties and digital preservation." https://web.archive.org/web/20160520082501/http:/www.significantproperties.org.uk/.

Significant Properties (archived version). "Testing Reports." https://web.archive.org/web/20160416031256/http://www.significantproperties.org.uk/testingreports.html.

Spreadsheets. "Modified EUSES Corpus." https://spreadsheets.ist.tugraz.at/index.php/corpora-for-benchmarking/euses/.

STATA. https://www.stata.com/.

Strodl, S. et al. "The Delos Testbed for Choosing a Digital Preservation Strategy." Springer. http://dx.doi.org/10.1007/11931584_35.

Veenendaal, R. Van, et al. "Significant Properties of Spreadsheets." https://doi.org/10.17605/OSF.IO/G8D5Y.

Veenendaal, R. Van, et al. “Significant Properties of Spreadsheets: An Update on the work of the Open Preservation Foundation’s Archives Interest Group.” iPRES 2019 – 16th International Conference on Digital Preservation.
Verdegem, R. & J. Slats. “Practical experiences of the Dutch digital preservation test-bed.” *VINE* 34, no. 2 (2004): 56-65.
Wijsman, L. “Catalogue Significant Properties of Spreadsheets.” Zenodo. https://doi.org/10.5281/zenodo.3902080 .
Wijsman, L. “The Significant Properties of Spreadsheets: Stakeholder Analysis.” Zenodo. https://doi.org/10.5281/zenodo.3971833.
Wikipedia. “Numeric precision in Microsoft Excel.” https://en.wikipedia.org/wiki/Numeric_precision_in_Microsoft_Excel.
Wikipedia. “Research Libraries Group.” https://en.wikipedia.org/wiki/Research_Libraries_Group.
Wilson, A. “Significant Properties Report.” https://significantproperties.kdl.kcl.ac.uk/wp22_significant_properties.pdf.
Zenodo. “The Significant Properties of Spreadsheets (OPF AIG Final Report).” https://doi.org/10.5281/zenodo.5387099.

Appendices

Appendix A: Characterisation Tools

FITS

Tool Name	File Information Tool Set (FITS)
Source URL	https://projects.iq.harvard.edu/fits/home
Description	"The File Information Tool Set (FITS) identifies, validates and extracts technical metadata for a wide range of file formats. It acts as a wrapper, invoking and managing the output from several other open-source tools. Output from these tools are converted into a common format, compared to one another and consolidated into a single XML output file."

Fido

Tool Name	Format Identification for Digital Objects (fido)
Source URL	https://openpreservation.org/products/fido/
Description	"Fido (Format Identification for Digital Objects) is an open-source command-line tool to identify the file formats of digital objects."

Siegfried

Tool Name	Siegfried
Source URL	https://github.com/richardlehane/siegfried
Description	"Siegfried is a signature-based file format identification tool, implementing: ▪ The National Archives UK's PRONOM file format signatures ▪ Freedesktop.org's MIME-info file format signatures ▪ The Library of Congress's FDD file format signatures (beta) ▪ Wikidata (beta)"

Lingfo (XLRD)

Tool Name	Lingfo, now XLRD
Source URL	https://xlrd.readthedocs.io/en/latest/index.html
Description	"Lingfo provides a library for developers to use to extract information from Microsoft Excel spreadsheet files. Versions of Excel supported: 2003, 2002, XP, 2000, 97, 95, 5.0, 4.0, 3.0. Support for Excel 2007 XML files is on the way." (COPTR, Page last updated on 11 June 2007) "Xlrd is a library for reading data and formatting information from Excel files in the historical .xls format."

Dependency Discovery Tool

Tool Name	Dependency Discovery Tool
Source URL	https://sourceforge.net/projects/officeddt/
Description	"The Dependency Discovery Tool searches through binary office files (.doc, .xls and .ppt) and tries to find any documents or files that are linked to the document."

Officeparser.py

Tool Name	Officeparser.py
Source URL	https://github.com/unixfreak0037/officeparser
Description	"Officerparser.py is a python script that parses the format of OLE compound documents used by Microsoft Office applications. Some useful features of this script include: ▪ Macro extraction ▪ Embedded file extraction ▪ Format analysis"

Ssconvert

Tool Name	Ssconvert
Source URL	https://github.com/paulfitz/gnumeric
Description	"Ssconvert is a command line utility to convert spreadsheet files between various spreadsheet file formats. It is a companion utility to Gnumeric, the powerful spreadsheet program created by the GNOME project."

Python-oletools

Tool Name	Python-oletools
Source URL	https://www.decalage.info/python/oletools
Description	"Python-oletools is a package of python tools to analyze Microsoft OLE2 files (also called Structured Storage, Compound File Binary Format or Compound Document File Format), such as Microsoft Office documents or Outlook messages, mainly for malware analysis, forensics and debugging."

Apache POIFS

Tool Name	Apache POI - POIFS
Source URL	https://poi.apache.org/components/poifs/
Description	"POIFS is a pure Java implementation of the OLE 2 Compound Document format."

Apache Tika

Tool Name	Apache Tika
Source URL	https://tika.apache.org/
Description	"The Apache Tika™ toolkit detects and extracts metadata and text from over a thousand different file types (such as PPT, XLS, and PDF)."

Python Libraries

Tool Name	Pywin32
Source URL	https://github.com/mhammond/pywin32
Description	"The Python for Win32 (pywin32) extension, which provides access to many of the Windows APIs from Python."
Notes	Used to create a python script that tells you if there is at least one hyperlink in a workbook (.xls file), at least one formula or at least one named object. Can be extended to e.g. used rows/columns and formatting.

Tool Name	Odfpy
Source URL	https://pypi.org/project/odfpy/
Description	"Odfpy is a library to read and write OpenDocument v. 1.2 files."
Notes	Odfpy should provide interfaces for Open Document Format, similar to pywin32.

Spreadsheet Complexity Analyser

Tool Name	Spreadsheet Complexity Analyser (SCA)
Source URL	https://github.com/RvanVeenendaal/Spreadsheet-Complexity-Analyser
Description	"This software extracts values of Excel spreadsheet properties and calculates a tentative spreadsheet complexity assessment based on threshold values."

Appendix B: Lists

List of properties

The list of properties or blue sheet is available as a separate spreadsheet: https://doi.org/10.5281/zenodo.5387099. Please note that we emptied the column AIG Person, as we found it irrelevant for our final result who was the first person to work on the investigation of a specific property.

Property lists

The table below show our initial property groups and properties from the NANETH Stakeholder Analysis (NANETH SA) and the Object Analysis (OA)

	Initial OA property groups	Initial NANETH SA property groups	Initial OA properties	Initial NANETH SA properties
1	Appearance	Application Settings	Caption Properties	1904 Date System
2	Application settings	Cell Content	Color Properties	3D Geometry
3	Auditing	Cell Formatting	Default Styles	3D Lighting
4	Cell content	Charts	Enhanced Graphic Styles	3D Material
5	Cell formatting	Comments	Graphic Styles	3D Picture Options
6	Comments	Data compression	Markup language	3D Shadow
7	Compression settings	Data Tools	Page Styles and Layout	3D Shapes Options
8	Context	Editing	Shadow Properties	3D Texture
9	Creation	External Data	Style Element	Accounting Format
10	Data format	Formatting	Styles	ActiveX Controls
11	Data sources	Formulas	Text Animation	Advanced Table Cells
12	Declaration	Graphic Elements	Header Footer Formatting	Advanced Table Model
13	Editing	Hyperlinks	Auto calculation	Advanced Tables
14	Form content	Localization	Automatic reload	Annotation
15	Formulas	Macros	Backgroup refresh	Area Chart
16	Integrity	Metadata	Fill Properties	Arranged Objects

	Initial OA property groups	Initial NANETH SA property groups	Initial OA properties	Initial NANETH SA properties
17	Localization	Pivot Tables	Master pages	Auditing Tracer Arrows
18	Macros	Printing	Worksheet row limit	Author
19	Objects	Protection	Changes to Excel source data	Auto Calculation
20	Page layout	Statistics	Has hyperlinks	Automatic Reload
21	Pivot tables	Tables	Hyperlink basis	Background
22	Printing		Hyperlink behaviour	Backgroup Refresh
23	Range		Hyperlinks changed	Banded Columns
24	Scenarios		Links up to date	Banded Rows
25	Security		Auditing tracer arrows	Bar Chart
26	Sharing		Change tracking	Basic Table Model
27	Statistics		Change Tracking Metadata	Basic Text Content
28	Summary		Customized error values and empty cell values	Body Element and Document Types
29	Table formatting		Data validation restrictions and messages	Border Formatting
30	TBD		Basic Text Content	Box and Whisker Chart
31	User agent		Lists	Bubble Chart
32	User definitions		Spreadsheet Document Content	Calculated Fields
33			Character and cell formatting	Calculated Items
34			Column Formatting Properties	Camera Tool/Paste as Picture Link Object
35			Conditional formatting	Caption
36			Font Face Declarations	Category

	Initial OA property groups	Initial NANETH SA property groups	Initial OA properties	Initial NANETH SA properties
37			Indented formats	Category Axis Title
38			Indented text	Category/Series Labels
39			Multiple fonts in a single cell	Cell Comments (or Notes)
40			Paragraph Formatting Properties	Cell Fill
41			Paragraphs and Basic Text Structure	Cell Inset Margin
42			Pattern fills	Cell References
43			Rotated or vertical text	Cell Styles
44			Text Alignment Properties	Cell Text Wrap
45			Text Fields	Cell Threaded Comments
46			Text Formatting Properties	Change Tracking
47			Text Styles	Change Tracking Metadata
48			Cell comments	Changes to Excel Source Data
49			Remarks	Character and Cell Formatting
50			Zip Bit Flag	Character Count
51			Zip Compressed Size	Character Set
52			Zip Compression	Chart Data Source
53			Zip CRC	Chart Layouts
54			Zip File Name	Chart Sheets
55			Zip Modify Date	Chart Styles
56			Zip Required Version	Chart Title
57			Zip Uncompressed Size	Code Page

	Initial OA property groups	Initial NANETH SA property groups	Initial OA properties	Initial NANETH SA properties
58			Author	Codes
59			Created	Color
60			Creating application name	Column Chart
61			Creating application version	Column Formatting
62			Creation date	Column Width
63			Initial creator	Combo Chart
64			Template	Company
65			Template	Conditional Format
66			1904 date system	Connector
67			Data Styles	Consolidation
68			Date format	Created
69			Format	Creating Application Name
70			Measure Properties	Creating Application Version
71			Number formats	Creation Date
72			Consolidation	Cube Functions
73			Data consolidation	Currency Format
74			Connector Properties	Custom Calculations
75			DDE Connections	Custom Format
76			External links	Custom Shapes
77			Relationships	Custom Sort Order
78			Table DDE Links	Custom Views

	Initial OA property groups	Initial NANETH SA property groups	Initial OA properties	Initial NANETH SA properties
79			Web queries	Customized Error Values and Empty Cell Values
80			Cell references	Data Labels
81			Last modifed by	Data Pilot Tables
82			Editing cycles	Data Styles
83			Editing duration	Data Tables
84			Last modifed by	Data Validation
85			Last modified	Data Validation Restrictions and Messages
86			Modified date	Database Functions
87			Total Edit Time	Database Ranges
88			Outlining and grouping	Date and Time Functions
89			Event Listener Tables	Date Format
90			Filters	Dates before 1900-01-01
91			Form Content	DDE Connections
92			Slicer	Default Styles
93			Calculated fields	Description
94			Calculated items	Document Security
95			Codes	Doughnut Chart
96			Custom calculations	Drawing Object Layers
97			Formulas	Drawing Shapes
98			Labels in formulas	Drop Lines
99			Subtotals	Editing Cycles

	Initial OA property groups	Initial NANETH SA property groups	Initial OA properties	Initial NANETH SA properties
100			Valid	Editing Duration
101			Wellformed	Embedded Objects
102			Character set	Encryption
103			Code page	Engineering Functions
104			Language	Enhanced Graphic Styles
105			Language	Error Bars
106			Thai alignment	Event Listener Tables
107			Macro sheet	Excel Form Controls
108			Macros	External Data Ranges
109			Scripts	External Hyperlinks
110			Visual Basic for Applications (VBA) projects	External Links
111			Annotation	File Name
112			3D Geometry Properties	File Permissions
113			3D Lighting Properties	Fill
114			3D Material Properties	Filter
115			3D Shadow Properties	Financial Functions
116			3D Shapes	First Column
117			3D Texture Properties	Floating Frame Formatting
118			Chart sheet	Font Face Declarations
119			Charts	Font Types
120			Custom Shapes	Form Content

	Initial OA property groups	Initial NANETH SA property groups	Initial OA properties	Initial NANETH SA properties
121			Drawing object layers	Format
122			Drawing Shapes	Format Version
123			Embedded objects	Formulas
124			Floating Frame Formatting Properties	Fraction Format
125			Frame Formatting Properties	Frame Formatting
126			Graphs	Frames/Borders
127			Has embedded objects	Frozen Panes
128			Inserted objects	Funnel Chart
129			Office Apps	General Format
130			Pivotcharts	Graphic Styles
131			Scale crop	Group and Outline
132			Sparklines	Grouped Items in Fields
133			Stroke Properties	Grouped Objects
134			Page Layout	Grouping
135			Page Layout Formatting Properties	Has Embedded Objects
136			Printing and page setup features	Header Footer Formatting
137			Grouped items in fields	Header Row
138			Data Pilot Tables	Header/Footer
139			pivot tables	Heading Pairs
140			PivotTable reports	Hide and Unhide Columns
141			Last printed	Hide and Unhide Rows

	Initial OA property groups	Initial NANETH SA property groups	Initial OA properties	Initial NANETH SA properties
142			Printed by	Hi-Low Lines
143			Database ranges	Histogram Chart
144			External data ranges	Horizontal Alignment in Cell
145			Scenarios	Hyperlink Basis
146			Document security	Hyperlink Behaviour
147			File Permissions	Hyperlink Formatting
148			Is protected	Image Border
149			Is rights managed	Image Effects
150			Password settings	IMBI PivotTables
151			Protection permissions	Indented Formats
152			Security	Indented Text
153			Share document	Information Functions
154			Shared Workbook information	Information Rights Management (IRM)
155			Character Count	Initial Creator
156			Document statistic	Ink Annotations
157			Number of Pages	Inserted Clip Art
158			Pagecount	Inserted Equations
159			Word Count	Inserted Image
160			Category	Inserted Objects
161			Company	Inserted Shapes
162			Description	Inserted Symbols
163			Document Metadata	Internal Hyperlinks

	Initial OA property groups	Initial NANETH SA property groups	Initial OA properties	Initial NANETH SA properties
164			File Name	Is Protected
165			Keyword	Is Rights Managed
166			Manager	Keyword
167			Metadata Elements	Labels in Formulas
168			MIME type	Language
169			Organization	Last Column
170			Size	Last Modified By
171			Subject	Last Modified
172			Title	Last Printed
173			Title Of Parts	Layout
174			Work process	Leader Lines on Data Labels
175			Advanced Table Cells	Legends
176			Advanced Table Model	Line Chart
177			Advanced Tables	Line Formatting
178			Basic Table Model	Links up to Date
179			Table Cell Formatting Properties	Lists
180			Table Formatting Properties	Locked Cell
181			Table Row Formatting Properties	Logical Functions
182			Table Styles	Lookup and Reference Functions
183			Table Templates	Macro Sheet
184			Body Element and Document Types	Macros

	Initial OA property groups	Initial NANETH SA property groups	Initial OA properties	Initial NANETH SA properties
185			Custom sort order	Manager
186			Custom views	Map Chart
187			Frames --> Borders	Margins
188			Heading Pairs	Markup Language
189			Page fields in rows or columns	Master Pages
190			Status	Math and Trigonometry Functions
191			Text Declarations	Measure
192			Producer	Merged Cells
193			User defined metadata	MIME Type
194			User-defined function categories	Modified Date
195			Format version	Multiple Fonts in a Single Cell
196			Version date	Names
197			Version log	Number Format
198			Versions	Number of Pages
199				Object Borders
200				Object Fills
201				Object Visibility
202				Objects in Charts
203				OLAP Formulas
204				OLAP Pivots
205				OLE Objects

	Initial OA property groups	Initial NANETH SA property groups	Initial OA properties	Initial NANETH SA properties
206				Organization
207				Outlining and Grouping
208				Page Breaks
209				Page Count
210				Page Fields in Rows or Columns
211				Page Layout
212				Page Layout Formatting
213				Page Orientation
214				Page Styles
215				Paragraphs and Basic Text Structure
216				Pareto Chart
217				Password Settings
218				Pattern Fills
219				Percentage Format
220				Picture Cropping
221				Picture Recoloring
222				Picture Styles
223				Pictures
224				Pie Chart
225				Pivot Tables
226				Pivot Table Reports

	Initial OA property groups	Initial NANETH SA property groups	Initial OA properties	Initial NANETH SA properties
227				Print Ranges
228				Printed By
229				Printing and Page Setup Features
230				Producer
231				Protected Sheet
232				Protected Workbook
233				Protection Permissions
234				Query Tables
235				Radar Chart
236				Regular Expressions (RegEx)
237				Relationships
238				Repeat Rows/Columns
239				Rich Text in Cell
240				Rotated or Vertical Text
241				Row Height
242				Row Heights/Columns Widths
243				Scale Crop
244				Scenarios
245				Scientific Format
246				Scripts
247				Security

	Initial OA property groups	Initial NANETH SA property groups	Initial OA properties	Initial NANETH SA properties
248				Series Axis Title
249				Series Data Source
250				Series Order
251				Shadow
252				Shape Styles
253				Shapes
254				Shapes on Charts
255				Share Document
256				Shared Workbook Information
257				Shared Workbooks
258				Sheet/Book Settings
259				Show Data Table
260				Show Legend Keys in Data Table
261				Show Series Major Gridline
262				Show Series Minor Gridline
263				Signature Line Object
264				Size
265				Slicers
266				SmartArt Diagrams
267				SmartArt Graphics
268				Sort

	Initial OA property groups	Initial NANETH SA property groups	Initial OA properties	Initial NANETH SA properties
269				Sort Table
270				Spark Lines
271				Special Format
272				Splits
273				Statistical Functions
274				Status
275				Stock Chart
276				Stroke Styles
277				Style Element
278				Styles
279				Subject
280				Subtotal
281				Sunburst Chart
282				Surface Chart
283				Table Cell Formatting
284				Table DDE Links
285				Table Formatting
286				Table Row Formatting
287				Table Styles
288				Table Templates
289				Template
290				Text Alignment

	Initial OA property groups	Initial NANETH SA property groups	Initial OA properties	Initial NANETH SA properties
291				Text Animation
292				Text Boxes
293				Text Declarations
294				Text Fields
295				Text Format
296				Text Functions
297				Text Styles
298				Thai Alignment
299				Time Format
300				Themes
301				Title
302				Title of Parts
303				Total Edit Time
304				Total Rows
305				Tracked Changes
306				Treemap Chart
307				Trendlines
308				User Defined Metadata
309				User-defined Function Categories
310				Valid
311				Value Axis Title

	Initial OA property groups	Initial NANETH SA property groups	Initial OA properties	Initial NANETH SA properties
312				Version Date
313				Version Log
314				Versions
315				Vertical Alignment in Cell
316				Visual Basic for Applications (VBA) Projects
317				Waterfall Chart
318				Web Queries
319				Well-formed
320				Window Settings
321				Word Count
322				WordArt
323				Work Process
324				Worksheet Row Limit
325				Worksheets
326				XY (Scatter) Chart
327				Zip Bit Flag
328				Zip Compressed File
329				Zip Compression
330				Zip CRC
331				Zip File Name
332				Zip Modify Date

	Initial OA property groups	Initial NANETH SA property groups	Initial OA properties	Initial NANETH SA properties
333				Zip Required Version
334				Zip Uncompressed Size

List of behaviours

The list of behaviours or green sheet is available as a separate spreadsheet: https://doi.org/10.5281/zenodo.5387099. Please note that we emptied the column AIG Person, as we found it irrelevant for our final result who was the first person to work on the investigation of a specific property.

List of specifications

The list of spreadsheet file format specifications and other information about spreadsheet file formats is included below.

List of specifications / publicly available information sources

- Apple Numbers
  - Native format: PUID: fmt/825, https://en.wikipedia.org/wiki/Numbers_(spreadsheet), .numbers files
- Gnumeric
  - Native format: PUID: fmt/1219, https://help.gnome.org/users/gnumeric/stable/gnumeric.html#file-format-gnumeric, .gnumeric/gnum/gnm, gzipped XML files, see also https://en.wikipedia.org/wiki/Gnumeric
- VisiCalc
  - Native format: PUID: x-fmt/368, http://fileformats.archiveteam.org/wiki/VisiCalc
  - Data interchange Format: PUID: x-fmt/41, http://fileformats.archiveteam.org/wiki/Data_Interchange_Format
- Lotus 1-2-3
  - Version 2: PUID: x-fmt/114, http://fileformats.archiveteam.org/wiki/Lotus_1-2-3, .wks/wk1/wk2/wk3/wk4/123 files
- Lotus Improv
  - Native format: PUID: n/a, https://en.wikipedia.org/wiki/Lotus_Improv, .imp files, see also https://fileinfo.com/extension/imp
- Quattro Pro
  - Spreadsheet for DOS, versions 1-4: PUID x-fmt/121, http://fileformats.archiveteam.org/wiki/WQ1, .wq1 files
  - Spreadsheet for DOS, versions 5.5, 5.5: PUID x-fmt/122, http://fileformats.archiveteam.org/wiki/WQ2, .wq2 files
  - Spreadsheet for Windows, versions 1-5: PUID fmt/834, http://fileformats.archiveteam.org/wiki/WB1, .wb1 files

    - Spreadsheet for Windows, version 6: PUID fmt/835, http://fileformats.archiveteam.org/wiki/WB2, .wb2 files
    - Spreadsheet, version 7,8: PUID fmt/836, http://fileformats.archiveteam.org/wiki/WB3, .wb3 files
    - Spreadsheet, version 9-12, X3, X4: PUID fmt/837, http://fileformats.archiveteam.org/wiki/QPW, .qpw files
    - See also http://fileformats.archiveteam.org/wiki/Quattro_Pro or the WordPerfect Office x7 handbook http://www.corel.com/static/product_content/wordperfect/x7/wpox7_user_guide_en.pdf
- Microsoft Excel
    - See https://en.wikipedia.org/wiki/Microsoft_Excel for information on the Microsoft Excel
    - Microsoft Office Excel 2003 (v 11.0);
        - Microsoft released in 2008 the specifications for Excel 2.0-11.0 https://msdn.microsoft.com/en-us/library/office/gg615597(v=office.14).aspx;
        - OpenOffice compiled their own documentation for the Excel format up to version 11: http://www.openoffice.org/sc/excelfileformat.pdf.
- OpenDocument Spreadsheet Document Format (ODS)
    - See https://en.wikipedia.org/wiki/OpenDocument for information on the Open Document Format and especially the Open Document Spreadsheet Format

Non-publicly available spreadsheet file formats

- Google Sheets
    - https://en.wikipedia.org/wiki/Google_Docs,_Sheets,_and_Slides

More information about the Excel Binary File Format

The Excel Binary File Format (.xls) Structure specifies the Excel Binary File Format (.xls). The Excel Binary File Format (.xls) is a collection of records and structures that specify workbook content, which can include unstructured or semi-structured tables of numbers, text, or both numbers and text, formulas, external data connections, charts, and images. Workbook content is typically organized in a grid-based layout, and often includes numeric data, structured data, and formulas.

More information about the Office Open XML SpreadsheetML file format and the Office Open XML file formats

- Office Open XML File Formats:
    - ISO/IEC 29500 (2008, 2011, 2012, 2016) consists of the following parts, under the general title Information technology — Document description and processing languages — Office Open XML File Formats:
        - Part 1: Fundamentals and Markup Language Reference
        - Office Open XML SpreadsheetML File Format
        - Part 2: Open Packaging Conventions
        - Part 3: Markup Compatibility and Extensibility
        - Part 4: Transitional Migration Features

- Microsoft's MSDN provides information on the Extensions to the Office Open XML SpreadsheetML File Format: https://msdn.microsoft.com/en-us/library/dd922181(v=office.12).aspx.

More information about the OpenDocument Spreadsheet Document Format and the Open Document Format

- Open Document Format
  - The content of ISO/IEC 26300-1 and OASIS OpenDocument v1.0 2nd ed. is identical.
    - ISO/IEC 26300-1 consists of the following parts, under the general title Information technology — Open Document Format for Office Applications (OpenDocument) v1.2:
      - Part 1: OpenDocument Schema
      - Part 2: Recalculated Formula (OpenFormula) Format
      - Part 3: Packages
  - Information about the Open Document Format (OpenDocument and OpenFormula) is available from https://www.oasis-open.org/standards#opendocumentv1.2.

Appendix C: Stakeholder questionnaire (sample by DNA)

Questions for data producers

Concerning users of the format

1. Which spreadsheet format do you use?
    a. If Excel, which version of Microsoft Office do you use?
2. Who in the organisation uses the format?
3. How many users have you approximately?
4. How often do you use the format? Multiple times daily, daily, weekly etc.

Concerning usage

5. What do you typically use the format for? E.g. casework, administration, budgets, project management, HR tasks, reporting, ad hoc tasks etc.
6. Is the chosen format vital for the usage?
7. Why did you choose this format instead of others?
8. Do you see alternative formats you could use? If not, why?
9. Which functions of the format do you use? If possible, prioritise the functions e.g. pivot charts, sorting and filtering, formulas, diagrams etc.

Concerning quantities and prevalence

10. How many spreadsheets do you assess are actively in use in your organisation? These can also be ranges e.g. “less than 100”, “100-1,000”, “1,000-10,000” etc.
11. Do you have an estimate on the size of your total number of spreadsheets? Could be in gigabyte or number of files of an avg. file size.
12. What is your assessment of the prevalence of the format within your use cases?
13. Do you share the data of the format with users outside of your organisation?
    a. If yes, do you export the data to other formats?

Concerning the future

14. Do you have areas today where you use spreadsheets, which in time you wish to use another format or application for?
15. Which measures do you in effect for securing the long-term preservation of spreadsheets? E.g. procedures, naming conventions, minimum criterias for the format, versioning etc.
16. Have you experienced not being able to open old spreadsheets?
17. How and what are your wishes for the submission and reuse of data sent to the Danish National Archives in the future?

Questions for archives

Concerning the archive

1. Brief presentation of the archive and your most important areas of work

Concerning quantities and prevalence

2. Are converted spreadsheets typically a part of your information packages?
3. How many information packages with converted spreadsheets do you have in your collections?
4. Do you have an estimate on how large a percentage spreadsheets typically constitute in your information packages? Ie. number of files
5. Which spreadsheet formats do you have experience with ingesting? ODS (Open Office), OOXML (Excel), other?

Concerning significant properties

6. What do you assess are significant properties to preserve in spreadsheets?
7. Do you consider the current preservation specification for spreadsheets, which are issued by the Danish National Archives, preserves the content of spreadsheets in an authentic and lossless manner?

Concerning submission

8. How often do you experience errors in conversion from a spreadsheet format to TIFF?
9. What types of errors do you typically experience in conversion of spreadsheets?
10. Do you possibly have an estimate on the additional costs currently related to conversion of spreadsheets?
11. Do you receive inquiries from data producers, suppliers or users concerning specific wishes for the submission of spreadsheets?
12. Do you receive copies of "preservation-worthy" spreadsheets in other formats than specified by the Danish National Archives (the TIFF format). If yes, which?
13. Do you receive "preservation-unworthy" spreadsheets (ie. because of independent preservation policy, retro digitisation or data of local historical importance) in other formats than TIFF? If yes, which?
14. If you receive spreadsheets (acc. to 12 and 13) do you validate the data? If yes, how?

Concerning reuse

15. Do you experience a general satisfaction with the users of TIFF'ed spreadsheets?
16. Do you receive inquiries in the dissemination and reuse of spreadsheets in original formats (e.g. Excel, ODS)?
17. Do you know of any behaviors which users demand when reusing spreadsheets?
18. Have you experienced finding spreadsheets in your collections, that you have not been able to reopen or where the conversion has changed the spreadsheet in such a way, that the spreadsheet could not be presented to a user credibly?

Concerning the future

19. Do you have ideas on other approaches for preserving spreadsheets?
20. How and what are your wishes for the submission and reuse of data in the future?
21. If you could give the Danish National Archives one recommendation, what would it be?

Appendix D: List of AIG colleagues

List of AIG colleagues who contributed and especially this work at any point in time. Thank you:

- **National Archives of Denmark**
  Anders Bo Nielsen
  Asbjørn Skødt
  Frederik Holmelund Kjærskov
  Jan Dalsten Sørensen
  Phillip Mike Tømmerholt

- **National Archives of Estonia**
  Kati Sein
  Koit Saarevet
  Lauri Leht

- **National Archives of the Netherlands**
  Remco van Veenendaal
  Jacob Takema
  Lotte Wijsman
  Margriet van Gorsel
  Pepijn Lucker

- **Open Preservation Foundation**
  Becky McGuinness
  Carl Wilson
  Charlotte Armstrong

- **Preservica**
  Jack O'Sullivan
  Jon Tilbury