University of Amsterdam

Report Internship Dual Period

MA Archival and Information Studies

The Significant Properties of Spreadsheets: Stakeholder Analysis

Author

Lotte Wijsman

University Supervisor

Dr. M. Karabinos

Lecturer

Prof.dr. C. Jeurgens

Supervisor NANETH

R. van Veenendaal

July 13, 2020

Contents

Introduction

According to the Open Archival Information System (OAIS), an archive is "an organization, that intends to preserve information for access and use by a Designated Community".[1] However, over the years, (digital) preservation is increasingly becoming more difficult because of the continuing growth of data and the fact that more information is created each day. The archivist's[2] main question of what to preserve is thus intensifying progressively. For digitally encoded information to be preserved, the information needs to remain understandable and usable.[3] However, preserving every aspect of the original digital object is not only costly, but also infeasible and undesirable at times. A selection of significant properties needs to be made to establish actions that focus on preserving those properties.[4] With significant properties we refer to: "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."[5]

For spreadsheets, ensuring the information that is needed to remain understandable and usable is not an easy feat.[6] The meaning of columns and cells and the formulas behind the outcome are oftentimes needed to make the information comprehensible. However, some of the context and meaning is given by the program it was created in. In the past, with the disappearance of some of these applications such as Lotus 1-2-3[7], a loss of information could occur. The object ceases to be understandable and usable and is therefore not properly preserved. Ensuring long-term accessibility whilst still preserving the significant properties of spreadsheets is a challenge in the current system. Selecting a suitable format will help

[1] "Reference Model For An Open Archival Information System (OAIS)," accessed February 20th, 2020, https://public.ccsds.org/Pubs/650x0m2.pdf, p. 19.

[2] Archivists select and keep documents, photographs, sounds recording, and other records that have enduring value as reliable memories of the past, and they help people find and understand the information they need in those records. From: Richard Pearce-Moses, "Identity and Diversity: What is an Archivist," *Archival Outlook* March/April 2006: 3. http://www.pearcemoses.info/papers/AO4.pdf.

[3] David Giaretta, *Advanced Digital Preservation* (Berlin, Heidelberg: Springer-Verlag, 2011), p. 1.

[4] A. Dappert and A. Farquhar, "Significance Is in the Eye of the Stakeholder," *Proceedings of the 13th European conference on Research and advanced technology for digital libraries (EDCL 2009)*: p. 297.

[5] A. Wilson, "Significant Properties Report," accessed March 17th, 2020, https://significantproperties.kdl.kcl.ac.uk/methodology.html, p. 8.

[6] A note of caution must be given here. Naturally, it is always best to save the original and everything that comes with it. However, this research focusses on events where this is not possible, and choices have to be made.

[7] "Lotus 1-2-3," Wikipedia, accessed July 4th, 2020, https://en.wikipedia.org/wiki/Lotus_1-2-3.

the preservation process. In order to do so, the significant properties of spreadsheets need to be established first.[8] This research was started in coordination with the Archives Interest Group (AIG) of the Open Preservation Foundation (OPF), whose members are seeing the increasing challenge of spreadsheet preservation.

To establish the significant properties of spreadsheets, two analyses should, according to the InSPECT[9] Framework Report, be executed. The first one is the Object Analysis, which will be described further in the theoretical framework.[10] This analysis investigated the structure of spreadsheets by making use of property extraction tools and studying various file format specifications. After extraction, the individual properties were put into property groups to create a better oversight of the function of the property itself.[11] Later, compatibility tables by Microsoft[12] and Apple[13] were added to the list. These helped to further standardize the terminology, in order to mitigate ambiguity in discussions on the matter.

With the list of properties in the property groups, the Stakeholder Analysis can be set up. This will be the focus of this study as well. As aforementioned, the archive preserves information for access and use by a designated community. This designated community consists of potential consumers who should be able to understand the information that is preserved by the archive.[14] These consumers have an interest and can therefore be considered to be stakeholders.[15] Performing the Stakeholder Analysis will answer the research question: which properties of spreadsheets are to be considered significant by stakeholders?

After determining what type of stakeholders were important, 16 stakeholders were found that wanted to take part in this research. This research will use several methods to discern

[8] 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*, p. 397.

[9] Investigating the Significant Properties of Electronic Content over Time.

[10] Gareth Knight, "InSPECT Framework Report," accessed March 13th, 2020, https://significantproperties.kdl.kcl.ac.uk/methodology.html.

[11] R. van Veenendaal et al., "Significant properties of spreadsheets," p. 397.

[12] "Differences between the OpenDocument Spreadsheet (ods.) format and the Excel for Windows (.xlsx) format," Microsoft, accessed June 15th, 2020, https://support.microsoft.com/en-ie/office/differences-between-the-opendocument-spreadsheet-ods-format-and-the-excel-for-windows-xlsx-format-3db958c8-e0ac-49a5-9965-2c2f8afbd960.

[13] "Document compatibility with Microsoft Office," Apple, accessed June 15th, 2020, https://www.apple.com/mac/numbers/compatibility/.

[14] OAIS, p. 21.

[15] OAIS, p. 29.

what stakeholders consider to be significant and why this is the case. An explorative study will at first help to get to know the background of the stakeholders. How do these stakeholders use spreadsheets and how do they qualify their knowledge? After doing this explorative survey, a formal study design will be applied. A catalogue with accompanying questions and interviews will help gain a further understanding of what is significant.

The explorative and formal study will complement each other and will result into several conclusions. For a study on the significant properties of spreadsheets, it is not beneficial to simply state several properties or property groups that must always be preserved. However, this research can provide archivists with what is generally significant and accompanying tools to help establish the significant properties per collection they have. This can even be extrapolated to other digital objects that archives have. Therefore, this research will have a contribution that is twofold. Firstly, this research will be the first to have executed a Stakeholder Analysis whilst considering the significant properties of spreadsheets. Secondly, it will give archivists and researchers tools and a methodology that can be used.

All in all, this study makes it evident that there not a singular suitable format to preserve spreadsheets. The chosen format relies heavily on the functionalities present in the spreadsheet. In general, in seems to be the case that individuals that have more knowledge of spreadsheets make use of more advanced functionalities. This study proposes a further refinement of the Spreadsheet Complexity Analyser as a potential tool to assess which format is suitable to the specific spreadsheet. This could provide an addition to preservation intent statements.

To reach feasible conclusions, a theoretical framework will first be discussed to elaborate on the work that has already been done regarding the significant properties of spreadsheets. After this, the next chapter will expound on how data will be collected and the methodology that will be used for this research. The final chapter will then expand on the results that were found. In the conclusion, these results will be discussed with further information on limitations of this research and a final recommendation.

1 Theoretical Framework

The study of significant properties was initially broadly researched by the preservation community. However, recently, very little attention from academia has been attributed towards this subject. As noted by Van Veenendaal et al., the subject is disappearing from the agenda of the preservation community.[16] Nonetheless, there is still a lot of untapped potential research to be done on the subject of significant properties. Exploring those pathways for further research has become of increasing urgency. The National Archives of the Netherlands (NANETH) are experiencing an expanding supply of spreadsheets that qualify for long term preservation.[17] This is due to article 12 of the Public Records Act 1995, where transfer of records happens after a set period of time.[18] Also, the Danish National Archives currently have six preferred formats for documents that are brought to their archives. None of these are spreadsheets formats. To assess the potential impact this may have on the retention of the important information of spreadsheets, research must be done towards significant properties of spreadsheets. As noted by Lynch, it is important to be able to guarantee that a reformatted version is the equivalent of an original document, with respect to specific properties.[19]

A problem that needs to be assessed concerns the terminology. Over the years, various terms have been used for significant properties. This made the subject ill-defined and difficult to grasp. Among the terms included were significant characteristics, significant properties, aspects, and essence.[20] In this study, the term significant characteristics is not used. A characteristic is made up out of a property and a value.

[16] R. van Veenendaal, P.C.M Lucker and C.D. Sijtsma, "Significant significant properties," accessed March 13th, 2020, https://openpreservation.org/wp-content/uploads/2018/10/Significant-Significant-Properties.pdf, p. 2.

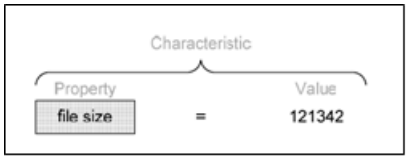
[17] "A period of time long enough for there to be concern about the impacts of changing technologies, including support for new media and data formats, and of a changing Designated Community, on the information being held in an OAIS." From: OAIS, p. 22.

[18] "Archiefwet 1995," Overheid.nl, accessed March 11th, 2020, https://wetten.overheid.nl/BWBR0007376/2020-01-01.

[19] Clifford A. Lynch, "Canonicalization: A Fundamental Tool to Facilitate Preservation and Management of Digital Information," *D-Lib Magazine* (September 1999), accessed June 20th, 2020, www.dlib.org/dlib/september99/09lynch.html.

[20] Dappert and Farquhar, "Significance," p. 298.

For example, the property being the file size and the value being the unit of measure (see figure). Because the significance is not taken per case, but in a broader sense, values are not included in this research project. The term significant properties is therefore more accurate and correct in this study.



The building blocks of a characteristic. From: Dappert and Farquhar, "Eye of the Stakeholder," p. 299.

This problem prevented researchers from reaching a consensus on which properties existed, let alone which were significant. In an attempt to overcome this hurdle, the AIG, also containing members from NANETH and the Danish National Archives, initiated a study. They decided to use the methodology from the InSPECT project.[21] In this methodology, two analyses are prescribed: the Object Analysis and the Stakeholder Analysis. The Object Analysis contained various steps that were performed by the AIG. Selecting spreadsheets as the object type and analysing the structure were the first two steps. In order to analyse the structure, a list of technical properties should be attained.[22] Numerous extractions tools, such as Apache Tika[23], Dependency Discovery Tool[24], and the New-Zealand Metadata Extraction Tool[25], were applied to make this list. Another method was to study various file format specifications.[26] However, these solutions did not account for the extraction of spreadsheet-specific properties, such as cells and formulas. In addition to this, the AIG wanted to aggregate spreadsheets into various types, such as simple/static and complex/dynamic. In the simple/static type fall the spreadsheets that simply contain rows and columns involving static values and simple formatting. On the other side are the complex/dynamic spreadsheets that contain macros and formulas for example.[27]

These two matters resulted in the creation of the Spreadsheet Complexity Analyser,

[21] Knight, "InSPECT Framework Report."
[22] Knight, "InSPECT Framework Report," 3.1.2.
[23] Apache Tika, accessed July 7th, 2020, https://tika.apache.org/.
[24] "Dependency Discovery Tool," SourceForge, accessed July 7th, 2020, https://sourceforge.net/projects/officeddt/.
[25] "Metadata Extraction Tool," SourceForge, accessed July 7th, 2020, http://meta-extractor.sourceforge.net/.
[26] R. van Veenendaal et al., "Significant properties of spreadsheets," p. 397.
[27] "Spreadsheet Complexity Analyser," GitHub, accessed June 10th, 2020, https://github.com/RvanVeenendaal/Spreadsheet-Complexity-Analyser.

which analyses the complexity of spreadsheets by extracting spreadsheets properties and their values. However, the aggregation of two types of spreadsheets was deemed too simplistic after one test put 99% of a 180.000 spreadsheet test set into the complex/dynamic type. Still, the tool filled a gap in "the property extraction and migration quality assessment tool ecosystem".[28]

The following step of the Object Analysis methodology was to classify the properties into several categories: content, context, rendering, structure, and behaviour.[29] Next to this, the properties were aggregated into self-made groups. This decision was made to find a middle ground between the specificity of individual properties and the complexity of the categories of the InSPECT methodology. In January 2020, the list containing the properties, categories and groups was further altered by making use of compatibility tables. Compatibility tables show the formatting differences between two file formats. For this study, the main importance of the compatibility table was the use of language. They mention areas, which the AIG have named property groups. Moreover, the compatibility tables give subareas, which are the properties themselves. For example, the group could be formatting, and the property could be row height/column widths. The compatibility tables from Apple[30] and Microsoft[31] are assumed to be closer to the use of language by stakeholders. They are therefore valuable to the Stakeholder Analysis.

This study is a continuation of the work by the AIG. The start of this study involved further refinement to the list of properties with accompanying categories and groups. The list was then used for the Stakeholder Analysis.

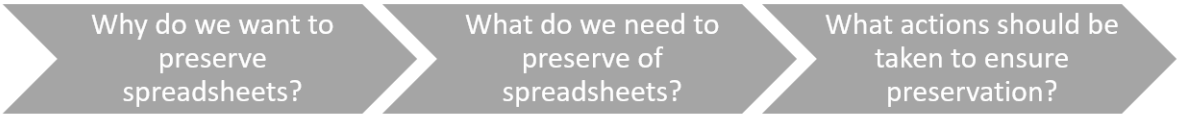
[28] R. van Veenendaal et al., "Significant properties of spreadsheets," p. 397.
[29] Knight, "InSPECT Framework Report." 3.1.3.
[30] Apple, "Document compatibility."
[31] Microsoft, "Differences."

2 Data and Methodology

2.1 Objectives

2.1.1 Preservation Intent

Preservation intent describes why and how we should strive to preserve certain digital objects, in this case spreadsheets.[32] A vital part of preservation intent is identifying which properties are to be considered significant. While this is the focus of this report, preservation intent is a broader concept. In the figure below, the three questions of preservation intent are visualised.



This research will focus mostly on the first two questions. It starts by asking the question: "Why are we doing this?".[33] Dutch governmental bodies are obliged to comply to the Public Records Act 1995. Article 3 states that governmental bodies must deliver and keep records held by them in good, orderly, and accessible condition.[34] Another reason to preserve can be to save time and effort, and increase efficiency. However, saved documents can have a goal as well. Spreadsheets are for example created to analyse data or to show departmental results. Herein lies a great difficulty because everyone uses the same piece of software to comply to all kinds of needs. In a seminal piece, Owens states that what to preserve, one must also have a clarified preservation intent. Users care about the meaning of the object, not the object itself. Not only having access to the object, but also keeping it accessible by preserving the meaning.[35] For this research, it is therefore not acceptable to just state what properties are considered significant, but it should also aim to deepen the knowledge of the

[32] "Essentiële kenmerken," Leren Preserveren, accessed May 29th, 2020, https://lerenpreserveren.nl/topic/essentiele-kenmerken/.

[33] Trevor Owens, *The Theory and Craft of Digital Preservation* (Baltimore: Johns Hopkins University Press, 2018), p. 82.

[34] Overheid.nl, "Archiefwet 1995."

[35] Owens, *Digital Preservation*, p. 82.

preservation intent and how this plays a role.

The National Library of Australia (NLA) considers the context of what must be achieved to be vital to create methods and solutions to go with them. By constructing 'preservation intent statements', the organisation's intentions are recorded for specific classes of digital content.[36] These statements encapsulate the intention behind the preservation and problems concerning the specific digital object.[37] For this research, having a single preservation statement made by an archive or library is, however, not sufficient. As aforementioned, spreadsheets have been used for a multitude of purposes. The National Archives of the Netherlands receives numerous spreadsheets coming from various public organisations and ministries. One preservation statement cannot suffice for all these spreadsheets coming from these diffused institutes.

Interviews can be used to determine motives behind the stakeholders arguments and choices. The interviews combined with the questions that were further asked, will answer the first two questions of preservation intent as presentation in the figure above, and will clear the way for research for the third question. Naturally, details can vary per collection and type of digital object but establishing patterns and creating a methodology can be of great use.

2.1.2 Property Aggregation

The Object Analysis cultivated into 334 properties in total. Many of these properties are very closely related, such as inserted shapes and inserted objects. It would be impossible to ask stakeholders to make a selection from all these. Therefore, adding a certain type of grouping is imperative. In the choice of the aggregation level, there lie two options.

The first option is provided by the InSPECT Framework Report. These are different categories of behaviours. Every property has a purpose that should be determined. These five categories are those purposes put broadly. The following five categories were used:

1. Content: information content within the spreadsheet. Examples of this are text and

[36]"Statements of Preservation Intent," National Library of Australia, accessed June 10th, 2020, https://www.nla.gov.au/content/statements-of-preservation-intent.

[37]"Preservation Intent – Australian Government Web Archive," National Library of Australia, accessed June 10th, 2020, https://www.nla.gov.au/content/preservation-intent-australian-government-web-archive.

images.

2. Context: this describes the environment in which the spreadsheet was created and that has an influence on its intended meaning. Examples are the initial creator and creation date.

3. Rendering: this has an influence on how the spreadsheets looks. Examples are font colour and font size.

4. Structure: this describes how two or more types of content are related to each other. Examples of this are auto calculation and cell references.

5. Behaviour: these are the properties that demonstrate how the content interacts with other stimuli. An example of this are hyperlinks.[38]

The second option is to subsume the individual properties into overarching groups. This work had already been started by the OPF AIG and was recently continued to further refine the groups. The difficulty here lies in finding a balance in specificity. On the one hand, having broader categories makes it easier to subsume the properties into. However, on the other hand, too broad can have detrimental effects to finding the significant properties. After refining the initial spreadsheet concerning the individual properties and the groups they were in, a new distribution was made. 21 groups came out of this:

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

[38]Knight, "InSPECT Framework Report." 3.1.3.

The largest group, with 15.6% of the properties, is graphic elements. The smallest group, with 0.3% of the properties, is cell content.

2.2 Setup

2.2.1 Identifying Stakeholders

2.2.1.1 Population

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. This community consists of stakeholders that should have no trouble in understanding information that the archive has preserved.[39] For this study, I sought out a population of individuals who were at least somewhat familiar with spreadsheets. This was necessary in order to ensure that the participating stakeholders were familiar with a broad range of proposed functionalities. However, their level of knowledge can differ. Gaining insights from multiple types of spreadsheet users is imperative to finding certain patterns. Also, their role may differ.

Three types of roles were set up prior to the start of the Stakeholder Analysis: 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 every group must have one or more representatives.[40] In this study, 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 organisation for which this study is carried out, the National Archives of the Netherlands, preserves information from public institutes. Moreover, this ensures that the spreadsheets that are considered in the research are similar to the spreadsheets that would eventually be preserved by national archives in the OPF AIG.

[39] OAIS, p. 21.
[40] Knight, "InSPECT Framework Report." 3.2.

2.2.1.2 Sample

Getting stakeholders to participate was a fairly difficult process due to the coronavirus. Several expansions of the search were needed to obtain 16 stakeholders to participate. The relationship managers of the National Archives reached out to several of their contacts within the ministries and other public organisations. After several weeks, this amounted to a number of eight stakeholders. After this, it was decided to extend the search. By posting a blog on KIA (Kennisnetwerk Informatie en Archief)[41] and also messaging several networks, eventually it was possible to get to a total of 16 stakeholders. Seven of those stakeholders are employed by various Dutch ministries, whilst the other nine are working for semi-governmental institutes. They range from policy advisor, to consultants and specialists. Of the stakeholders, four are female and eight are male. Their knowledge of spreadsheets is diverse because of their diversity in function. Because of the fairly small sample, it may prove to be more difficult to find statistically significant results.

2.2.1.3 Explorative Study

To understand the background of the stakeholders, they were first sent some explorative questions. The primary purpose of these questions was to obtain background information, but they also helped to form important questions for the formal study design later. The five questions asked were the following:

- What is your full name and function?
- What do you create or use spreadsheets for?
- How do you qualify your knowledge of spreadsheets (the use of different functionalities)?
- What is your opinion about which five properties of a spreadsheet are important if the spreadsheet must be preserved?

[41] "Onderzoek essentiële kenmerken van spreadsheets," Kennisnetwerk Informatie en Archief, Lotte Wijsman, accessed May 11th, 2020, https://kia.pleio.nl/groups/view/28319672/kennisplatform-informatiehuishouding-overheden/blog/view/55814242/onderzoek-essentiele-kenmerken-van-spreadsheets.

- In which of the following roles do you see yourself and could you give some explanation?

The first three questions are purely to start with some simple questions that already give a lot of information about the stakeholder. A person that uses spreadsheets for difficult analyses can make very different choices in the catalogue later on then someone that merely uses it for information storage. Knowledge also comes into to play here. A study by Lawson et al. researched how experienced spreadsheet users compared to those less experienced. In addition to that, the study also looked at how these different groups made use of spreadsheets. More experienced users are more aware of spreadsheet risk and this leads towards a higher devotion toward spreadsheet design.[42] A higher level of experience/knowledge can also lead towards a higher use frequency of certain functionalities, such as macros and pivot tables.[43]

Asking the stakeholders which five properties of a spreadsheet are important if the spreadsheet must be preserved also indicates their preferences and background. Interestingly, this question was interpreted in two ways. Some interpreted the question as what five properties needed to be preserved if it were decided that it would be preserved. Others interpreted the question as which properties could help with the preservation process. An example of these two interpretations can be taken from one property given by stakeholder 4 and 5.

- Interpretation 1: retention of formulas/calculations (stakeholder 4).
- Interpretation 2: use of meaningful column and row titles (stakeholder 5).

Collaborating with this question was also the fifth question about role. A manager can have a completely different view and expertise than a maker. The InSPECT Framework Report claims that a clear understanding is needed "of the relationship between the stakeholder that is the target of analysis and the object type".[44] It is simply not feasible to have a 'on size fits all' solution for the significant properties of spreadsheets. Therefore, intentions and background are extremely important to come to any conclusions and/or patterns.

Furthermore, the stakeholders were asked to submit a spreadsheet that is representative. The spreadsheet must be characteristic for the spreadsheets that they use daily in their

[42]Barry R. Lawson, Kenneth R. Baker, Stephen G. Powell & Lynn Foster-Johnson, "A comparison of spreadsheet users with different levels of experience," *Omega* 37, no. 3 (2009): p. 584-585.
[43]Lawson et al., "spreadsheet users," p. 587.
[44]Knight, "InSPECT Framework Report." 3.2.

functions. Because all the stakeholders are from the public sector, it might occur that information is confidential and can therefore not be shared. The participants were also allowed to submit an already publicly available spreadsheet. 11 of the 16 stakeholders submitted their spreadsheets.

The reason for utilizing spreadsheets in this study is twofold. The first usage is to visually see how the stakeholders make use of spreadsheets and if the five properties they have selected (question 4 of the explorative study) can possibly be related to this. Moreover, the choices made later on could be explained through the submitted spreadsheet. If a stakeholder works with formulas on a daily basis, it is only natural that these are to be considered significant by that person. The second usage is data driven. A Spreadsheet Complexity Analyser was made by the AIG to extract information from spreadsheets. Both of these usages focus on identifying whether the properties that were selected by the stakeholder are in line with the properties that are used in their respective spreadsheets.

2.2.2 Formal Study Design

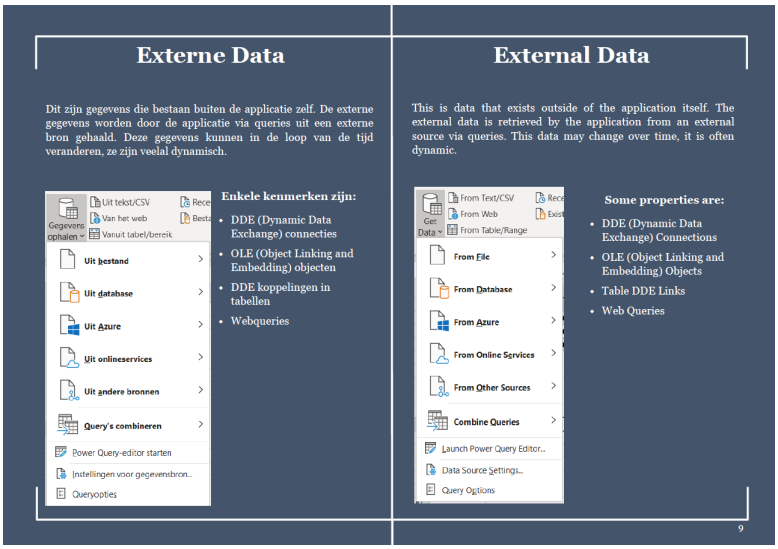
2.2.2.1 Catalogue

The first part of the formal study design was constructed out of the Object Analysis. This analysis generated 334 properties. To let stakeholders pick out of this very large number would not be feasible and therefore groups were further created. At the beginning of this study, further work was done in order to refine this overview. At the end of this work, 21 groups were created. The biggest group (Graphic Elements) contains 52 properties, while the smallest group (Cell Content) contains just 1 property. With this grouping, a note of caution must be added. Some properties, such as 'Objects in Charts', can easily fit into two groups (Graphic Elements or Charts). Effort was put into fitting the properties in the groups they would fit in the best. This resulted in the allocation as is seen in Appendix A.

To further clarify the groups, the idea of the catalogue was born.[45] The catalogue is meant to portray the 21 groups in an orderly fashion. Two important choices had to be made during the creation process. The first choice was language. As mentioned in the introduction, this

[45]"Catalogue Significant Properties of Spreadsheets (June 20,2020)," Zenodo,Lotte Wijsman, accessed June 20th, 2020, http://doi.org/10.5281/zenodo.3902080.

study wants to add a toolbox and methodology to the field of significant properties and stakeholder analyses. By making use of not only the Dutch language, but also the English language, international comparability is enhanced. These reasonings led to the catalogue being in two languages. Every page is in Dutch and English with a clear border between the two languages. The second choice had to be made between what could be deemed as application-specific property groups. Application-specific properties are properties that are embedded into specific spreadsheet software, such as Microsoft Excel or Numbers by Apple. The properties are not part of the spreadsheet itself but can still be important enough to be preserved. Four groups were considered to be application-specific: data compression, localization, printing and application settings. In the catalogue, they are divided from the other groups to clearly indicate that they exist outside of the spreadsheets. Every property group is explained through a small explanation, visual examples, and examples of properties inside that group.



Page 9 of the catalogue

Stakeholders were sent this catalogue with the question to pick out the five groups they considered to be most significant. They were also asked to motivate their answers and to enter which application they worked with most when using spreadsheets. The five groups of properties they had to select were connected to the explorative study, where they also

had to select five properties that they deemed significant. However, the choices in the explorative study were unrestricted, whilst the catalogue had fixed groups to choose from. Some stakeholders indicated that it was hard to make choices because some groups cohere. There was also the challenge that some property groups are only significant when they are used, such as macros, pivot tables and charts.

2.2.2.2 Formal Interviews

From the beginning, it was clear that interviews were imperative for this research. However, a conscious decision was made to have the interviews be the last part of the Stakeholder Analysis. If the interviews had been done first, a bias could have been included from the part of the evaluator (myself). Without any background information, the questions would also have been too general. By having access to the answers from the explorative study, having received the representative spreadsheet, and having the answers about the catalogue, the questions were more specific. Five in-depth interviews were carried out during this final phase. The stakeholders were selected on the initial time it took them to complete the explorative survey and the catalogue. The interviews were semi-structured. A set list of questions was asked, as well as impromptu questions about their answers given during the interview itself.

The set list of questions[46] was made up of two parts. The first part was about the first step of preservation intent: Why do we want to preserve spreadsheets? By asking why the stakeholder as an individual wants to preserve certain spreadsheets, and why the organisation has an interest in preserving spreadsheets, it can be combined with the introductory question of why the stakeholders make or use spreadsheets. Together, these make up the first step of preservation intent. Another important feature is the current policy. Does their organisation have a set policy in place when it comes to the preservation of spreadsheets? Where are the spreadsheets saved? Are they saved locally or in an overarching system? Among these questions was also a question relating back to the what needs to be preserved, the significant properties. Every time, the stakeholder makes a conscious choice to make use of a spreadsheet application instead of a word processor. What drives this decision and is this related to any

[46]See Appendix D

functionalities or properties? This leads to the second part of the questions.

The second part of the set interview questions was about the use of the spreadsheet. Here three options are present. The first option is storage, where the spreadsheet is used as a simple database containing numbers or text. This is without formatting and pictures. The second option is static representation. This is still plain text or numbers, but with formatting and potential images. This can still be converted to a PDF file without encountering loss. The third option is dynamic representation. These are the spreadsheets involving 'dynamic' functions, such as formulas, external data, and links. Converting these to a PDF can simply not suffice and will result in a considerable loss of data.

The interviews will not only bring about the preservation intent for these individuals, but will also lead to a better understanding of what needs to be preserved.

3 Results

3.1 Qualitative Results

16 stakeholders in total participated in this study. 11 of those also handed in representative spreadsheets. As in the explorative study, the first non-demographic question was about why they make and/or use spreadsheets. Here, several things came to the fore. Seven stakeholders clearly mentioned using spreadsheets for their reports and analyses. However, the most stated rationale was to create an overview. Stakeholders are using spreadsheets to visualize, to use filters or pivot tables.

Two questions of the explorative survey were categorised, the questions where stakeholders qualified their knowledge level and stated their role. The first was not set up beforehand into stated categories, but after the answers came in it was clear that all stakeholders themselves used a similar system of qualification, namely basic/average/advanced. Two stakeholders said that their knowledge was basic, five stated to have an average knowledge, and nine stated an advanced knowledge level. Naturally, they categorised themselves and it can be that what someone perceives as basic, another person views as average. Basic and average could be put together here, which makes the distribution seven for basic/average and nine for advanced. It will be considered whether the results of the statistical analyses are robust to treating these two subgroups as one entity. In the statistical analysis, these two categorisations are further explored. The second question, about the role of the stakeholders, was a requirement of the Stakeholder Analysis methodology of the InSPECT Framework Report. Three roles were defined for this research: the maker, user, and manager. The role of the maker was predominantly chosen (11 times), with two stakeholders filling in maker and user. The stakeholder analysis showed that these two roles are often intertwined and therefore it was decided to fuse these together as a group. In this case, 14 stakeholders fall into this group, with just two stakeholders falling into the manager group.

The stakeholders also filled in the reply form of the catalogue. An overview of this can be found in Appendix C. Four groups have not been selected by any stakeholders: statistics, data compression, localization, and printing. The five most selected groups were:

- Formulas (chosen 13 times). A formula calculates the value of a cell (or multiple cells). For example. It can calculate 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 summarizes 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.

Formulas was the group chosen the most. One of the three stakeholders that did not opt for this group, already gave a reasoning for this in the explorative study. When asked for a qualification of their level of knowledge, this stakeholder indicated to have an average level of knowledge. An outline of what is possible concerning the functionalities in Excel was known, but the stakeholder would not mess around with formulas. Looking into the spreadsheet that was submitted by this stakeholder, this was indeed confirmed. This spreadsheet was the most sizeable of the spreadsheets that were submitted, with 2.115.462 cells used, but contained 0 formulas. The other two stakeholders that did not opt for this group, did not have clear motives to not opt for this. Stakeholder number 4 even mentioned formulas in the five properties of the explorative study, but then decided to instead opt for cell content, cell formatting, external data, charts, and pivot tables in the catalogue. It can be concluded from this, that one size does not fit all. Dappert and Farquhar already stated this in their

paper, that significance is not binary, nor absolute. We can only discern patterns and use these to establish the significant properties for a certain stakeholder.[47]

Reasonings given by stakeholders to opt for this property group, were because formulas enrich data and were often used for analyses. Some already considered the preservation process in their motivation and stated that formulas tell you about how a spreadsheet is built and is working. Missing the calculations behind the outcome therefore constitutes a substantial loss of data for these stakeholders. Interestingly, during the explorative study, the stakeholders were asked in free form what five properties they considered significant in a spreadsheet. Ten stakeholders already had formulas mentioned in those. Of these ten, nine eventually chose the property group formulas in the catalogue.

Four of the five groups that were chosen the most to be significant, are dynamic groups. Formulas, external data, pivot tables and charts are part of the dynamic representation group mentioned in the setup of the formal interviews. As mentioned before, there were four application-specific property groups: application settings, printing, data compression, and localization. Of these, only application settings was chosen once. Stakeholder number 13 was the only one choosing application settings as a significant property group, reasoning that settings strongly determine how the worksheet functions.

The interviews further confirmed the information provided by the explorative study and the catalogue. The stakeholders, when asked why they opted for spreadsheets instead of text documents such as Word, all stated functionality was their main reason. Formulas was the most prominent named functionality as well, next to external links. This is in line with the choices made in the catalogue. Coinciding with this is their reason to preserve. They use the previously made spreadsheet as a way to look back and to remember. To keep the content understandable, the functionalities they deem significant need to be preserved. This is of course highly dependent on the functionality used by the stakeholders themselves. All in all, one can conclude that very specific functionality in certain context is of crucial importance. The interviews therefore have provided further evidence to support these intitial notions that came forth from the explorative study. It appears that of the three options (storage, static representation and dynamic representation), dynamic representation is not only the most

[47] Dappert and Farquhar, "Significance," p. 303.

option to preserve, but is also the most necessary option.

3.2 Statistical Analysis

3.2.1 Explorative Survey and Catalogue

Other than simply letting stakeholders pick out the five groups they considered to be most significant,[48] I further wanted to explore if the answers given in the exploratory study could have an influence in the choices that were made. Using STATA[49], several analyses were made. First, four analyses concerning the stakeholders level of knowledge were executed. The first analysis that was performed concerned gender and knowledge. The study by Lawson et al. on spreadsheet users included demographic parameters such as gender. Group A, who had a lower level of capability, and group B, who were deemed the most advanced. In their research, group A tended to be more populated by females.[50] Of the 16 stakeholders in my study, four were female. The third question in the explorative study asked the stakeholders to qualify their level of knowledge. The answers given could be brought back to basic, average, and advanced. By tabulating a two-way table setting gender against knowledge, table 1 was produced. Also, a Pearson $\chi 2$ test was added.

Gender	Knowledge			
	Basic	Average	Advanced	Total
Male	0	4	8	12
	0.0%	33.3%	66.7%	100%
Female	2	1	1	4
	50%	25%	25%	100%
Total	2	5	9	16
	12.5%	31.3%	56.3%	100%
Probability			0.030	

Table 1: Tabulation of knowledge and gender

[48] See Appendix C.

[49] STATA is software for statistics and data science. STATA, accessed April 3rd, 2020, https://www.stata.com/.

[50] Lawson et al., "spreadsheet users," p. 582-583.

This statistical test can be applied to data in order to evaluate the probability that an observed difference between groups arose by happenstance. The null hypothesis in this test is that there are no systematic differences between the groups.[51] With a probability of 0.03, there is a 97% chance that the groups indeed differ in terms of their knowledge. This is in line with the paper by Lawson et al. However, when we treat the subgroups basic and average as one entity, the result becomes statistically insignificant (probability of 0.146). Important here is to note that the level of knowledge was self-reported. It could be the case that some stakeholders have under- or overestimated themselves.

Another analysis that can be done using the stakeholders level of knowledge is concerning the property group pivot tables. Lawson et al. researched how often the two types of groups made use of certain Excel features. Group A, with a lower level of capability, reported an infrequent use of pivot tables, while group B, with an advanced level of capability, reported occasional to frequent use.[52] With a higher relative frequency of use, a higher level of knowledge can often be involved.

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

Table 2: Tabulation of the property group pivot tables and knowledge

Table 2 shows how level of knowledge might influence if people chose pivot tables as one of the five significant property groups. With a 96.3% chance it shows that a higher level of knowledge indeed leads towards a higher percentage of choosing the property group pivot

[51] James T. McClave, P. George Benson and Terry Sincich, *Statistics for Business and Economics* (London: Pearson, 2017), p. 603.

[52] Lawson et al., "spreadsheet users," p. 587.

tables. This result is robust to treating basic and average knowledge as one (probability of 0.01).

The third analysis tabulated the knowledge of the stakeholder against the version they used. In Appendix C the results of the catalogue are shared, among those which spreadsheet applications the stakeholders work the most with, and which version of it. The overarching applications already given as options were Excel, Numbers and Google Sheets. Furthermore, concerning Excel, several options were given as to which version they made use of (e.g. Excel 2010/2013/2016), and the stakeholders could add their own options as well. All stakeholders made use of Excel, with a majority (56.3%) making use of Excel 2016.

Knowledge	Version				
	Excel 2010	Excel 2013	Excel 2016	Microsoft 365	Total
Basic	1	0	0	1	2
	50%	0.0%	0.0%	50%	100%
Average	2	1	2	0	5
	40%	20%	40%	0.0%	100%
Advanced	0	1	7	1	9
	0.0%	11.1%	77.8%	11.1%	100%
Total	3	2	9	2	16
	18.8%	12.5%	56.3%	12.5%	100%
		Probability		0.152	

Table 3: Tabulation of application version and knowledge

However, as seen in table 3, a higher level of knowledge does not determine a newer Excel version. With a probability of 0.152, this is not statistically significant by traditional levels of significance. Even when treating basic and average as one, the result remains insignificant (probability of 0.132)

The final analysis regarding the stakeholders level of knowledge concerned the role they identify with: maker, user or manager. It is not unreasonable to believe that a maker has a higher level of knowledge concerning the functionalities of spreadsheet application than a manager. However, in practice it appeared that the roles of maker and user were extremely difficult to separate. Stakeholders oftentimes felt that they belonged to both groups. Therefore, a change was made considering the roles. The maker and user groups are

merged together to form the maker/user group.

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	

Table 4: Tabulation of knowledge and role

Table 4 indeed shows 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.

After these four analyses concerning the level of knowledge, it was time to look at the second option mentioned in chapter 1: the five categories. The five categories from the InSPECT Framework Report were content, context, rendering, structure, and behaviour. When working on the Object Analysis, not only was every property put into a group, it was also put in a category.[53] In a spreadsheet, for every property it was stated in what property group it subsumed, as well as in what category. Then these two types of options were tabulated against each other, so a percentage comes out. This shows which percentage of properties in that property group belong to a certain category. This table can be found in Appendix E. Some of the properties in one group all belonged to the same category. However, some groups fell into different categories with an overruling percentage. At this point, choices needed to be made manually. Pivot tables, for example, fell into the categories rendering (50%) and structure (40%). Eventually, after looking at the individual properties, it was deemed that the lower percentage accompanying structure was a better fit here.

[53]See Appendix A.

After this analysis, it can be established how the stakeholders choose concerning these five categories. Are the property groups they have chosen pointing towards a significant category? In Appendix F we can see that rendering (14) and structure (15) are the categories that were chosen the most. However, the categories content, context, and behaviour were respectively chosen 9, 10, and 12 times. For this, the same limitation as mentioned earlier applies. With a larger sample size, bigger differences in significance might be encountered. What we can conclude from this is that the stakeholders choices are very diverse and fall into different categories, while using the groups yield fruitful results that show clear preferences.

3.2.2 Spreadsheet Complexity Analyser

As mentioned in the setup of the explorative study, every stakeholder was asked to submit a spreadsheet that was representative of what they encounter daily in their work. 11 of the 16 stakeholders submitted their spreadsheets. In order to analyse these, the Spreadsheet Complexity Analyser was used.[54] Because of the fact that the property group formulas was chosen the most by the stakeholders, a calculation was done to determine which percentage of the cells used was used for formulas ((formulas/cells used)*100 =%).[55] The final percentages varied from 0.0% to 54.6%, with an average of 22.2%. Also important were the number of fonts used and cell styles. These give an indication of the formatting and can therefore be correlated.

The outcomes of the Spreadsheet Complexity Analyser were subsequently linked to the results of the explorative study and the catalogue. Based on these data, various associations were scrutinised. Firstly, the percentage of formulas in cells used was correlated to the level of knowledge. A positive correlation of 0.15 was found, which leads us to attach weight to the fact that individuals possessing a higher level of knowledge are more prone to an increasing use of formulas. The result was robust to treating the knowledge levels basic and average as one; a positive correlation of 0.09 was found.

Secondly, the number of formulas used is found to have a positive association with the

[54] See Appendix G for the individual results.

[55] One stakeholder, stakeholder number 14, submitted three spreadsheets, so an average of those three spreadsheets was taken.

likelihood of an individual selecting the formulas group in the catalogue. The positive correlation between these two variables was of the magnitude 0.41. This indicates that when an individual uses more formulas (in percentage of cells used), we would be inclined to predict that the individual is also more likely to choose formulas as one of the significant property groups.

The third association is twofold. The Spreadsheet Complexity Analyser also distinguishes how many fonts or cell styles are present in a spreadsheet. The question attached to this was: is there a positive association between how many fonts or cell styles are used and if the individual chooses a formatting group as significant? Between fonts and formatting, a positive correlation was found of 0.24. Cell styles and formatting also presented a positive correlation of 0.38. This indicates that individuals that make use of more cell styles are also likely to attach significance to the property group formatting.

Discussion

Conclusion

This study aimed to answer the question: which properties of spreadsheets are to be considered significant by stakeholders? The Stakeholder Analysis assessed the stakeholders by making use of an explorative and formal study. With this, a methodology and toolbox were postulated for further research regarding the Stakeholder Analysis of the significant properties of spreadsheets.

There are two ways to assess the data that was collected. The first assessment was to look at the way the properties are aggregated. The groups that came forth as being the most significant from the formal study design of the catalogue were mostly dynamic functionalities, such as formulas, external data, and pivot tables. However, these are only significant when used. One can therefore conclude that one size does not fit all when considering the various ways spreadsheets can be preserved. For simple/static spreadsheets, converting to a PDF could suffice. However, when more advanced functionalities are utilised, other formats should be explored. One could think about databases and statistical programs, but the final choice should depend on the functionalities imbedded in the actual spreadsheet. Databases and statistical programmes were also suggested by the stakeholders in the interviews when asked what they would swerve to if spreadsheets ceased to exist. I therefore deem various formats to be necessary in order to be able to preserve a significant fraction of all considered spreadsheets.

The functionalities used are dependent on the level of knowledge of the individual. An individual with a less advanced knowledge level is less likely to use advanced functionalities and will therefore not consider them to be significant. As shown, the probability of stakeholders choosing the property group pivot tables was higher with an increased level of knowledge.

The second aggregation option was the five categories. These indicate the purpose each property has. By putting every property in a category as well as a group, it could then be calculated which percentage of the properties in a group fell into a certain category (see

Appendix E). If a stakeholder picked the property group cell content as significant, this was linked to the category of content. However, the results (as seen in Appendix F) were not conclusive concerning the categories.

The second assessment concerned the three options people can use spreadsheets as: storage, static representation, or dynamic representation. This was a further refinement of the tentative assessment that the AIG made for their Spreadsheet Complexity Analyser: simple/static and complex/dynamic. One of the questions of the interview was why people made the conscious choice to make a spreadsheet instead of a simple text document. Unanimously the stakeholders stated that the functionalities of Excel in their case were the reason for this. Formulas, using external data and the sort and filter options made spreadsheets the more attractive option. Making use of spreadsheets creates a clearer overview of the information and the options are more flexible. These fall within the option of the dynamic representation. This implicates that converting spreadsheets to PDF would not suffice in the case of preservation.

Preservation intent is often what drives people towards their choices. They create a spreadsheet to make, for example, analyses. For these analyses, they often make use of the pivot table functionality and therefore consider this to be significant. The preservation intent is why we need to preserve. This can not be just one intent, every stakeholder has a different use for spreadsheets. However, we can state that the original intention of the spreadsheet needs to be preserved. To do this, functionalities that the stakeholder deems to be of significance, need to be preserved. The interviews confirmed that some groups are chosen more often and can therefore be seen as significant in general. When approaching stakeholders in the future, these patterns can be used a starting point.

The statistical analysis determined certain patterns and probabilities. After seeing some patterns arise, such as the level of knowledge and the choice for pivot tables, an even further assessment was made by including the Spreadsheet Complexity Analyser. The SCA allows us to assess spreadsheets at face-value. It can directly be seen what type of functionalities are used by the stakeholder. Interestingly enough, the functionalities used in the representative spreadsheet concur with the chosen significant property groups. This indicates that the significant properties can be identified by making use of the characteristics of the spreadsheets

itself. The SCA could therefore function as an indicative tool to pre-appraise the appropriate format to which the spreadsheet should be converted.

Recommendations

Determining what the significant properties of spreadsheets are is not an easy feat. One size does not fit all. There is still room for more research here, using a more diverse and bigger sample size. A suggestion is to assess spreadsheets on face-value using the Spreadsheet Complexity Analyser. If this tool could be expanded to include pivot tables and external data, spreadsheets could be assessed from the start to fit a certain type as the AIG has suggested previously. Based on the outcome of the Spreadsheet Complexity Analyser, archives could make a qualitative assessment of which format would retain the maximum amount of functionality in the original spreadsheet. This leads to the possibility of using different formats without making the chosen format dependent solely on the merits of the preservation intent of the individual. This could provide an alternative to the sole use of a preservation intent statement. This could also be used as a check to confirm whether the preservation intent statement matches the actual spreadsheet.

Furthermore, I suggest an in-depth discussion on the methodology employed in order to identify the significant properties of spreadsheets. The toolbox and methodology that were created for this study were ad hoc. Discussion will give room to further refinement where this is necessary. This could provide further guidance to researchers and archivists that aim to identify the significant properties.

Limitations

As with any research, limitations were present. The main one being the sample size. Due to the coronavirus, interest to participate was low. This led to less data being available to study. Having a greater sample size could have benefitted the statistical analysis to find more significant relationships. Diversity was also a limitation. A more sizeable sample might have included more women and most of all, more stakeholders from the manager role. The effect

of role on the choices that were made is still untapped and further research must be done here. The sample was also self-selected, which can lead to having more experienced users instead of a diverse group concerning knowledge level. It could be the case that experienced users are not representative of the total population of all makers, users, and managers of spreadsheets. This would induce a bias to the results.

Another limitation was the lack of guidance from the literature. A study such as this has not been published before. This led to the development of a new methodology, rather than making use of a strategy that has been widely used before. The only guidance on how to execute a Stakeholder Analysis concerning significant properties came from the InSPECT Framework Report. However, this was very theoretical in nature, providing little practical guidance for applied research on this topic. Nonetheless, various procedures that were proposed by the InSPECT Framework Report were applied to this study, such as the categories and the role of the stakeholder.

Moreover, the lack of a clear-cut methodology led to the role of the researcher to be at the forefront. Several times, choices had to be made based on personal preferences. The aggregation of the properties into the groups and categories is a prominent example of this. One could argue that a degree of subjectivity is introduced here by the researcher. The in-depth discussion, mentioned in the recommendation section, could refine this where necessary eventually.

Bibliography

Apache Tika. Accessed July 7th, 2020. https://tika.apache.org/.

Apple. "Document compatibility with Microsoft Office." Accessed June 15th. 2020. https://www.apple.com/mac/numbers/compatibility/.

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)*: 297-308.

Giaretta, David. *Advanced Digital Preservation.* Berlin, Heidelberg: Springer Verlag, 2011.

Github. "Spreadsheet Complexity Analyser." Accessed June 10th. 2020. https://github.com/RvanVeenendaal/Spreadsheet-Complexity-Analyser.

Kennisnetwerk Informatie en Archief. "Onderzoek essentiële kenmerken van spreadsheets." Lotte Wijsman. Accessed May 11th, 2020. https://kia.pleio.nl/groups/view/28319672/kennisplatforminformatiehuishouding- overheden/blog/view/55814242/onderzoek-essentiele-kenmerken-van-spreadsheets.

Knight, Gareth. "InSPECT Framework Report."Accessed March 13th, 2020. https://significantproperties.kdl.kcl.ac.uk/methodology.html

Lawson, Barry R., Kenneth R. Baker, Stephen G. Powell & Lynn Foster-Johnson. "A comparison of spreadsheet users with different levels of experience." *Omega* 37, no. 3 (2009): 579-590.

Leren Preserveren. "Essentiële kenmerken." Accessed May 29th, 2020. https://lerenpreserveren.nl/topic/essentiele-kenmerken/.

Lynch, Clifford A. "Canonicalization: A Fundamental Tool to Facilitate Preservation and Management of Digital Information." *D-Lib Magazine* (September 1999). Accessed June 20th, 2020. www.dlib.org/dlib/september99/09lynch.html.

McClave, James T., P. George Benson & Terry Sincich. *Statistics for Business and Economics.* London: Pearson, 2017.

Microsoft. "Differences between the OpenDocument Spreadsheet (ods.) format and the Excel for Windows (.xlsx) format."Accessed June 15th, 2020. https://support.microsoft.com/

en-ie/oce/dierencesbetween- the-opendocument-spreadsheet-ods-format-and-the-excel-for -windows-xlsx-format-3db958c8-e0ac- 49a5-9965-2c2f8afbd960.

National Library of Australia. "Preservation Intent - Australian Government Web Archive." Accessed June 10th, 2020. https://www.nla.gov.au/content/preservation-intent-australian-government-web-archive.

National Library of Australia. "Statements of Preservation Intent." Accessed June 10th, 2020. https://www.nla.gov.au/content/statements-of-preservation-intent.

Overheid.nl. "Archiefwet 1995." Accessed March 11th, 2020. https://wetten.overheid.nl/BWBR0007376/2020-01-01.

Owens, Trevor. *The Theory and Craft of Digital Preservation*. Baltimore: John Hopkins University Press, 2018.

Pearce-Moses, Richard. "Identity and Diversity: What is an Archivist." *Archival Outlook* (March/April 2006): 3. http://www.pearcemoses.info/papers/AO4.pdf.

"Reference Model For An Open Archival Information System (OAIS).". Accessed February 20th, 2020. https://public.ccsds.org/Pubs/650x0m2.pdf.

SourceForge. "Metadata Extraction Tool." Accessed July 7th, 2020. http://metaextractor. sourceforge.net/.

SourceForge. "Dependency Discovery Tool." Accessed July 7th, 2020. https://sourceforge.net/projects/oceddt/.

STATA. Accessed April 3rd, 2020. https://www.stata.com/.

Van Veenendaal, R. et al. "Significant Properties of Spreadsheets: An update on the work of the Open Preservation Foundation's Archives Interest Group." *iPres 2019 - 16th International Conference on Digital Preservation*: 396 - 398.

Van Veenendaal, R, P.C.M. Lucker & C.D. Sijtsma. "Significant significant properties." Accessed March 13th, 2020. https://openpreservation.org/wp-content/uploads/2018/10/Significant-Significant-Properties.pdf.

Wikipedia. "Lotus 1-2-3." Accessed July 4th, 2020. https://en.wikipedia.org/wiki/Lotus_1-2-3.

Wilson, A. "Significant Properties Report."Accessed March 17th, 2020. https://significantproperties.kdl.kcl.ac.uk/methodology.html.

Zenodo. "Catalogue Significant Properties of Spreadsheets (June 20, 2020)." Lotte Wijsman. Accessed June 20th, 2020. June 20th, 2020, http://doi.org/10.5281/zenodo.3902080.

Appendix A: Properties, Groups and Categories

	Property	Group	Category
1	1904 Date System	Application Settings	Rendering
2	3D Geometry	Graphic Elements	Rendering
3	3D Lighting	Graphic Elements	Rendering
4	3D Material	Graphic Elements	Rendering
5	3D Picture Options	Graphic Elements	Rendering
6	3D Shadow	Graphic Elements	Rendering
7	3D Shapes Options	Graphic Elements	Rendering
8	3D Texture	Graphic Elements	Rendering
9	Accounting Format	Cell Formatting	Rendering
10	ActiveX Controls	Graphic Elements	Rendering
11	Advanced Table Cells	Tables	Structure
12	Advanced Table Model	Tables	Structure
13	Advanced Tables	Tables	Rendering
14	Annotation	Charts	Rendering
15	Area Chart	Charts	Rendering
16	Arranged Objects	Graphic Elements	Rendering
17	Auditing Tracer Arrows	Formulas	Structure
18	Author	Metadata	Context
19	Auto Calculation	Application Settings	Structure
20	Automatic Reload	Application Settings	Behaviour
21	Background	Graphic Elements	Rendering
22	Backgroup Refresh	Application Settings	Behaviour
23	Banded Columns	Tables	Rendering
24	Banded Rows	Tables	Rendering
25	Bar Chart	Charts	Rendering
26	Basic Table Model	Tables	Structure
27	Basic Text Content	Cell Content	Content
28	Body Element and Document Types	Metadata	Context
29	Border Formatting	Formatting	Rendering
30	Box and Whisker Chart	Charts	Rendering
31	Bubble Chart	Charts	Rendering
32	Calculated Fields	Pivot Tables	Content
33	Calculated Items	Formulas	Content
34	Camera Tool/Paste as Picture Link Object	Graphic Elements	Rendering
35	Caption	Formatting	Rendering
36	Category	Metadata	Context
37	Category Axis Title	Charts	Content
38	Category/Series Labels	Charts	Content
39	Cell Comments (or Notes)	Comments	Context
40	Cell Fill	Cell Formatting	Rendering

	Property	Group	Category
41	Cell Inset Margin	Formatting	Rendering
42	Cell References	Data Tools	Structure
43	Cell Styles	Cell Formatting	Rendering
44	Cell Text Wrap	Cell Formatting	Rendering
45	Cell Threaded Comments	Comments	Context
46	Change Tracking	Editing	Content
47	Change Tracking Metadata	Editing	Context
48	Changes to Excel Source Data	Application Settings	Behaviour
49	Character and Cell Formatting	Cell Formatting	Rendering
50	Character Count	Statistics	Content
51	Character Set	Localization	Rendering
52	Chart Data Source	Charts	Content
53	Chart Layouts	Charts	Rendering
54	Chart Sheets	Charts	Content
55	Chart Styles	Charts	Rendering
56	Chart Title	Charts	Content
57	Code Page	Localization	Rendering
58	Codes	Formulas	Context
59	Color	Formatting	Rendering
60	Column Chart	Charts	Rendering
61	Column Formatting	Formatting	Rendering
62	Column Width	Tables	Rendering
63	Combo Chart	Charts	Rendering
64	Company	Metadata	Context
65	Conditional Format	Cell Formatting	Rendering
66	Connector	External Data	Behaviour
67	Consolidation	Data Tools	Structure
68	Created	Metadata	Context
69	Creating Application Name	Metadata	Context
70	Creating Application Version	Metadata	Context
71	Creation Date	Metadata	Context
72	Cube Functions	Formulas	Behaviour
73	Currency Format	Cell Formatting	Rendering
74	Custom Calculations	Formulas	Structure
75	Custom Format	Cell Formatting	Rendering
76	Custom Shapes	Graphic Elements	Rendering
77	Custom Sort Order	Application Settings	Rendering
78	Custom Views	Application Settings	Rendering
79	Customized Error Values and Empty Cell Values	Formulas	Rendering
80	Data Labels	Charts	Content

	Property	Group	Category
81	Data Pilot Tables	Pivot Tables	Structure
82	Data Styles	Cell Formatting	Rendering
83	Data Tables	Charts	Content
84	Data Validation	Data Tools	Structure
85	Data Validation Restrictions and Messages	Data Tools	Structure
86	Database Functions	Formulas	Behaviour
87	Database Ranges	Data Tools	Structure
88	Date and Time Functions	Formulas	Structure
89	Date Format	Cell Formatting	Rendering
90	Dates before 1900-01-01	Application Settings	Rendering
91	DDE Connections	External Data	Behaviour
92	Default Styles	Formatting	Rendering
93	Description	Metadata	Context
94	Document Security	Protection	Context
95	Doughnut Chart	Charts	Rendering
96	Drawing Object Layers	Graphic Elements	Structure
97	Drawing Shapes	Graphic Elements	Rendering
98	Drop Lines	Charts	Rendering
99	Editing Cycles	Editing	Context
100	Editing Duration	Editing	Context
101	Embedded Objects	Graphic Elements	Rendering
102	Encryption	Protection	Context
103	Engineering Functions	Formulas	Structure
104	Enhanced Graphic Styles	Graphic Elements	Rendering
105	Error Bars	Charts	Rendering
106	Event Listener Tables	Graphic Elements	Structure
107	Excel Form Controls	Graphic Elements	Rendering
108	External Data Ranges	External Data	Behaviour
109	External Hyperlinks	Hyperlinks	Behaviour
110	External links	External Data	Behaviour
111	File Name	Metadata	Context
112	File Permissions	Protection	Context
113	Fill	Charts	Rendering
114	Filter	Data Tools	Structure
115	Financial Functions	Formulas	Structure
116	First Column	Tables	Rendering
117	Floating Frame Formatting	Formatting	Rendering
118	Font Face Declarations	Formatting	Rendering
119	Font Types	Formatting	Rendering
120	Form Content	Data Tools	Rendering

	Property	Group	Category
121	Format	Cell Formatting	Rendering
122	Format Version	Metadata	Context
123	Formulas	Formulas	Structure
124	Fraction Format	Cell Formatting	Rendering
125	Frame Formatting	Formatting	Rendering
126	Frames/Borders	Formatting	Rendering
127	Frozen Panes	Application Settings	Rendering
128	Funnel Chart	Charts	Rendering
129	General Format	Cell Formatting	Rendering
130	Graphic Styles	Graphic Elements	Rendering
131	Group and Outline	Data Tools	Rendering
132	Grouped Items in Fields	Pivot Tables	Rendering
133	Grouped Objects	Graphic Elements	Rendering
134	Grouping	Pivot Tables	Rendering
135	Has Embedded Objects	Metadata	Structure
136	Header Footer Formatting	Formatting	Rendering
137	Header Row	Tables	Rendering
138	Header/Footer	Printing	Rendering
139	Heading Pairs	Metadata	Structure
140	Hide and Unhide Columns	Formatting	Rendering
141	Hide and Unhide Rows	Formatting	Rendering
142	Hi-Low Lines	Charts	Rendering
143	Histogram Chart	Charts	Rendering
144	Horizontal Alignment in Cell	Cell Formatting	Rendering
145	Hyperlink Basis	Metadata	Behaviour
146	Hyperlink Behaviour	Hyperlinks	Behaviour
147	Hyperlink Formatting	Hyperlinks	Rendering
148	Image Border	Graphic Elements	Rendering
149	Image Effects	Graphic Elements	Rendering
150	IMBI PivotTables	Pivot Tables	Rendering
151	Indented Formats	Cell Formatting	Rendering
152	Indented Text	Cell Formatting	Rendering
153	Information Functions	Formulas	Structure
154	Information Rights Management (IRM)	Protection	Context
155	Initial Creator	Metadata	Context
156	Ink Annotations	Graphic Elements	Rendering
157	Inserted Clip Art	Graphic Elements	Rendering
158	Inserted Equations	Graphic Elements	Rendering
159	Inserted Image	Graphic Elements	Rendering
160	Inserted Objects	Graphic Elements	Rendering

	Property	Group	Category
161	Inserted Shapes	Graphic Elements	Rendering
162	Inserted Symbols	Graphic Elements	Rendering
163	Internal Hyperlinks	Hyperlinks	Structure
164	Is Protected	Protection	Context
165	Is Rights Managed	Protection	Context
166	Keyword	Metadata	Context
167	Labels in Formulas	Formulas	Content
168	Language	Localization	Context
169	Last Column	Tables	Rendering
170	Last Modified By	Editing	Context
171	Last Modified	Editing	Context
172	Last Printed	Printing	Context
173	Layout	Pivot Tables	Rendering
174	Leader Lines on Data Labels	Charts	Rendering
175	Legends	Charts	Rendering
176	Line Chart	Charts	Rendering
177	Line Formatting	Graphic Elements	Rendering
178	Links up to Date	Application Settings	Behaviour
179	Lists	Formatting	Rendering
180	Locked Cell	Protection	Rendering
181	Logical Functions	Formulas	Structure
182	Lookup and Reference Functions	Formulas	Structure
183	Macro Sheet	Macros	Behaviour
184	Macros	Macros	Behaviour
185	Manager	Metadata	Context
186	Map Chart	Charts	Rendering
187	Margins	Printing	Rendering
188	Markup Language	Formatting	Rendering
189	Master Pages	Application Settings	Structure
190	Math and Trigonometry Functions	Formulas	Structure
191	Measure	Cell Formatting	Rendering
192	Merged Cells	Tables	Rendering
193	MIME Type	Metadata	Context
194	Modified Date	Editing	Context
195	Multiple Fonts in a Single Cell	Cell Formatting	Rendering
196	Names	Formulas	Content
197	Number Format	Cell Formatting	Rendering
198	Number of Pages	Statistics	Content
199	Object Borders	Graphic Elements	Rendering
200	Objects Fills	Graphic Elements	Rendering

	Property	Group	Category
201	Object Visibility	Graphic Elements	Rendering
202	Objects in Charts	Graphic Elements	Rendering
203	OLAP Formulas	Formulas	Structure
204	OLAP Pivots	Pivot Tables	Structure
205	OLE Objects	External Data	Behaviour
206	Organization	Metadata	Context
207	Outlining and Grouping	Graphic Elements	Rendering
208	Page Breaks	Printing	Rendering
209	Page Count	Statistics	Content
210	Page Fields in Rows or Columns	Pivot Tables	Rendering
211	Page Layout	Printing	Rendering
212	Page Layout Formatting	Printing	Rendering
213	Page Orientation	Printing	Rendering
214	Page Styles	Printing	Rendering
215	Paragraphs and Basic Text Structure	Cell Formatting	Rendering
216	Pareto Chart	Charts	Rendering
217	Password Settings	Protection	Context
218	Pattern Fills	Cell Formatting	Rendering
219	Percentage Format	Cell Formatting	Structure
220	Picture Cropping	Graphic Elements	Rendering
221	Picture Recoloring	Graphic Elements	Rendering
222	Picture Styles	Graphic Elements	Rendering
223	Pictures	Graphic Elements	Rendering
224	Pie Chart	Charts	Rendering
225	Pivot Tables	Pivot Tables	Structure
226	Pivot Table Reports	Pivot Tables	Structure
227	Print Ranges	Printing	Rendering
228	Printed By	Printing	Context
229	Printing and Page Setup Features	Printing	Rendering
230	Producer	Metadata	Context
231	Protected Sheet	Protection	Context
232	Protected Workbook	Protection	Context
233	Protection Permissions	Protection	Context
234	Query Tables	External Data	Behaviour
235	Radar Chart	Charts	Context
236	Regular Expressions (RegEx)	Formulas	Structure
237	Relationships	External Data	Behaviour
238	Repeat Rows/Columns	Printing	Rendering
239	Rich Text in Cell	Cell Formatting	Rendering
240	Rotated or Vertical Text	Cell Formatting	Rendering

	Property	Group	Category
241	Row Height	Tables	Rendering
242	Row Heights/Column Widths	Formatting	Rendering
243	Scale Crop	Graphic Elements	Rendering
244	Scenarios	Data Tools	Structure
245	Scientific Format	Cell Formatting	Rendering
246	Scripts	Macros	Behaviour
247	Security	Protection	Context
248	Series Axis Title	Charts	Content
249	Series Data Source	Charts	Content
250	Series Order	Charts	Rendering
251	Shadow	Graphic Elements	Rendering
252	Shape Styles	Graphic Elements	Rendering
253	Shapes	Graphic Elements	Rendering
254	Shapes on Charts	Charts	Rendering
255	Share Document	Metadata	Behaviour
256	Shared Workbook Information	Metadata	Behaviour
257	Shared Workbooks	Metadata	Behaviour
258	Sheet/Book Settings	Application Settings	Context
259	Show Data Table	Charts	Rendering
260	Show Legend Keys in Data Table	Charts	Rendering
261	Show Series Major Gridline	Charts	Rendering
262	Show Series Minor Gridline	Charts	Rendering
263	Signature Line Object	Graphic Elements	Rendering
264	Size	Metadata	Context
265	Slicers	Data Tools	Structure
266	SmartArt Diagrams	Graphic Elements	Rendering
267	SmartArt Graphics	Graphic Elements	Rendering
268	Sort	Data Tools	Rendering
269	Sort Table	Data Tools	Rendering
270	Spark Lines	Charts	Structure
271	Special Format	Cell Formatting	Rendering
272	Splits	Application Settings	Rendering
273	Statistical Functions	Formulas	Structure
274	Status	Metadata	Context
275	Stock Chart	Charts	Rendering
276	Stroke Styles	Formatting	Rendering
277	Style Element	Graphic Elements	Rendering
278	Styles	Graphic Elements	Rendering
279	Subject	Metadata	Context
280	Subtotal	Formulas	Structure

	Property	Group	Category
281	Sunburst Chart	Charts	Rendering
282	Surface Chart	Charts	Rendering
283	Table Cell Formatting	Tables	Rendering
284	Table DDE Links	External Data	Behaviour
285	Table Formatting	Tables	Rendering
286	Table Row Formatting	Tables	Rendering
287	Table Styles	Tables	Rendering
288	Table Templates	Tables	Rendering
289	Template	Metadata	Context
290	Text Alignment	Cell Formatting	Rendering
291	Text Animation	Graphic Elements	Rendering
292	Text Boxes	Graphic Elements	Content
293	Text Declarations	Metadata	Structure
294	Text Fields	Cell Formatting	Content
295	Text Format	Cell Formatting	Rendering
296	Text Functions	Formulas	Structure
297	Text Styles	Cell Formatting	Rendering
298	Thai Alignment	Localization	Context
299	Time Format	Cell Formatting	Structure
300	Themes	Graphic Elements	Rendering
301	Title	Metadata	Context
302	Title of Parts	Metadata	Context
303	Total Edit Time	Editing	Context
304	Total Rows	Tables	Rendering
305	Tracked Changes	Editing	Content
306	Treemap Chart	Charts	Rendering
307	Trendlines	Charts	Rendering
308	User Defined Metadata	Metadata	Context
309	User-defined Function categories	Metadata	Context
310	Valid	Metadata	Context
311	Value Axis Title	Charts	Content
312	Version Date	Metadata	Context
313	Version Log	Metadata	Context
314	Versions	Metadata	Context
315	Vertical Alignment in Cell	Cell Formatting	Rendering
316	Visual Basic for Applications (VBA) Projects	Macros	Behaviour
317	Waterfall Chart	Charts	Rendering
318	Web Queries	External Data	Behaviour
319	Well-formed	Metadata	Context
320	Window Settings	Application Settings	Context

	Property	Group	Category
321	Word Count	Statistics	Content
322	WordArt	Graphic Elements	Rendering
323	Work Process	Metadata	Context
324	Worksheet Row Limit	Application Settings	Context
325	Worksheets	Statistics	Content
326	XY (Scatter) Chart	Charts	Rendering
327	ZIP Bit Flag	Compression Settings	Context
328	ZIP Compressed File	Compression Settings	Context
329	ZIP Compression	Compression Settings	Context
330	ZIP CRC	Compression Settings	Context
331	ZIP File Name	Compression Settings	Context
332	ZIP Modify Date	Compression Settings	Context
333	ZIP Required Version	Compression Settings	Context
334	ZIP Uncompressed Size	Compression Settings	Context

Appendix B: Profiles of the Stakeholders

In this table, the three levels of knowledge still apply. However, for further research, the categories basic and average should be considered as one.The role are already brought back to two categories: maker/users and managers.

	Stakeholder			
	1	2	3	4
Institute	Semi-governmental	Semi-governmental	Governmental	Governmental
Gender	Male	Male	Male	Male
Knowledge	Advanced	Advanced	Advanced	Advanced
Role	Maker/user	Maker/user	Maker/user	Maker/user
Application	Excel 2016	Excel 2013	Excel 2016	Excel 2016

	Stakeholder			
	5	6	7	8
Institute	Governmental	Semi-governmental	Semi-governmental	Governmental
Gender	Male	Female	Male	Female
Knowledge	Average	Average	Average	Basic
Role	Maker/user	Maker/user	Maker/user	Maker/user
Application	Excel 2010	Excel 2010	Excel 2013	Excel 2010

	Stakeholder			
	9	10	11	12
Institute	Semi-governmental	Governmental	Governmental	Semi-governmental
Gender	Male	Male	Male	Male
Knowledge	Advanced	Advanced	Average	Average
Role	Maker/user	Maker/user	Manager	Maker/user
Application	Excel 2016	Excel 2016	Excel 2016	Excel 2016

	Stakeholder			
	13	14	15	16
Institute	Semi-governmental	Semi-governmental	Governmental	Semi-governmental
Gender	Male	Female	Male	Female
Knowledge	Advanced	Advanced	Advanced	Basic
Role	Maker/user	Maker/user	Maker/user	Manager
Application	Excel 2016	Microsoft 365	Excel 2016	Microsoft 365

Appendix C: Results Catalogue

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			

Appendix D: Fixed Questions Interviews

Preservation Intent

1. Why is it useful for the organisation to keep spreadsheets?
2. Why is it useful for yourself to keep spreadsheets?
3. Is there a general policy in the organisation on how spreadsheets should be kept or does everyone decide this for themselves?
   (a) If organisation → explain.
   (b) If everyone for themselves → do you recognise differences between yourself and your colleagues?
4. How do you store the spreadsheets? Do you do this locally or in a system?
   (a) Do you save it as an .xlsx or .xls file or in a different format (e.g. PDF)?
5. How long will the spreadsheets be stored within your organisation?
6. Who has insight into these spreadsheets?
   (a) What about the privacy concerning these spreadsheets?
7. Why do you use spreadsheets instead of word processors (e.g. Word)?

Storage vs. Static Representation vs. Dynamic Representation

1. Who do you create the spreadsheets for and do you share them?
   (a) What do these possible external individuals do with it?
2. Do you use Excel to store raw data (numbers/names) before tables are created/analysed?
   (a) If not → where do you store this raw data and what about the use of external data?

(b) If yes → do you ever run into limitations of Excel?

3. Are the data (e.g. tables, charts, and results of formulas) ever converted to another form of representation (e.g. PDF or PowerPoint)?

   (a) If not → Why do you prefer Excel?

   (b) If yes → What do you think is better about this than Excel?

4. What would you do if Excel, and other spreadsheet applications, were to disappear? What would be the consequences for the already existing spreadsheets and what would you swerve to?

5. Do you create your own Excel templates so that you don;t have to re-create them later but only have to fill in the raw data?

Appendix E: Groups and Categories in Percentages

Groups	Categories				
	Content	Context	Rendering	Structure	Behaviour
Protection	0.0%	**91.7%**	8.3%	0.0%	0.0%
Editing	22.2%	**77.8%**	0.0%	0.0%	0.0%
Cell Content	**100%**	0.0%	0.0%	0.0%	0.0%
Cell Formatting	3.1%	0.0%	**90.6%**	6.3%	0.0%
Data Tools	0.0%	0.0%	33.3%	**66.7%**	0.0%
Pivot Tables	10%	0.0%	50%	**40%**	0.0%
External Data	0.0%	0.0%	0.0%	0.0%	**100%**
Formulas	13.6%	4.6%	4.6%	**68.2%**	9.1%
Charts	21.3%	2.1%	**74.5%**	2.1%	0.0%
Graphic Elements	1.9%	0.0%	**92.3%**	3.9%	1.9%
Hyperlinks	0.0%	0.0%	25%	25%	**50%**
Macros	0.0%	0.0%	0.0%	0.0%	**100%**
Metadata	0.0%	**81.6%**	0.0%	7.9%	10.5%
Formatting	0.0%	0.0%	**100%**	0.0%	0.0%
Comments	0.0%	**100%**	0.0%	0.0%	0.0%
Statistics	**100%**	0.0%	0.0%	0.0%	0.0%
Tables	0.0%	0.0%	**83.3%**	16.7%	0.0%
Data Compression	0.0%	**100%**	0.0%	0.0%	0.0%
Localization	0.0%	**50%**	50%	0.0%	0.0%
Printing	0.0%	16.7%	**83.3%**	0.0%	0.0%
Application Settings	0.0%	20%	**40%**	13.3%	26.7%

Appendix F: Categories Chosen by Stakeholders

Stakeholder	Categories				
	Content	Context	Rendering	Structure	Behaviour
1	x	x	x	x	
2		x	x	x	
3			x	x	x
4	x		x	x	x
5		x	x	x	x
6			x	x	x
7	x	x	x	x	x
8		x	x	x	x
9	x		x	x	x
10	x	x		x	x
11	x	x	x	x	x
12	x	x	x	x	
13		x	x		x
14	x		x	x	x
15			x	x	x
16	x	x		x	
Total	9	10	14	15	12

Appendix G: Results Spreadsheet Complexity Analyser

	Stakeholder		
	1	2	3
Size (in kB)	17	1420	40
Worksheets	1	7	2
Fonts	8	33	22
Defined names	1	77	0
Cell styles	20	227	17
Formulas	81	9300	152
Hyperlinks	0	0	0
Comments	1	2	0
VBA macros	0	0	0
Shapes	0	0	4
Dates	23	612	0
Cells used	588	281433	481
External Links	0	4	0
Revision history	0	0	0
Tentative assessment	Complex/Dynamic	Complex/Dynamic	Complex/Dynamic
Percentage formulas	13.8%	3.3%	31.6%

	Stakeholder		
	5	6	7
Size (in kB)	11136	39	166
Worksheets	13	5	4
Fonts	40	21	5
Defined names	10	1	0
Cell styles	231	14	35
Formulas	0	628	100
Hyperlinks	8	0	0
Comments	0	0	0
VBA macros	0	0	0
Shapes	0	1	0
Dates	0	0	0
Cells used	2115462	1259	350
External Links	0	0	0
Revision history	0	0	0
Tentative assessment	Complex/Dynamic	Complex/Dynamic	Complex/Dynamic
Percentage formulas	0.0%	49.9%	28.6%

	Stakeholder		
	8	11	13
Size (in kB)	306	22	567
Worksheets	22	4	6
Fonts	61	4	32
Defined names	27	0	17
Cell styles	375	76	168
Formulas	1306	121	267
Hyperlinks	0	0	1
Comments	2	0	1
VBA macros	0	0	9
Shapes	2	0	30
Dates	0	0	3
Cells used	9914	1127	1535
External Links	6	0	0
Revision history	0	0	0
Tentative assessment	Complex/Dynamic	Complex/Dynamic	Complex/Dynamic
Percentage formulas	13.2%	10.7%	17.4%

	Stakeholder		
	14a	14b	14c
Size (in kB)	15	26	234
Worksheets	2	4	1
Fonts	6	12	5
Defined names	3	3	2
Cell styles	8	36	37
Formulas	16	104	11482
Hyperlinks	0	0	0
Comments	0	0	0
VBA macros	0	0	0
Shapes	0	0	0
Dates	0	0	1311
Cells used	134	491	21032
External Links	0	0	0
Revision history	0	0	0
Tentative assessment	Complex/Dynamic	Complex/Dynamic	Complex/Dynamic
Percentage formulas	11.9%	21.2%	54.6%

	Stakeholder
	15
Size (in kB)	3952
Worksheets	3
Fonts	72
Defined names	2
Cell styles	276
Formulas	95631
Hyperlinks	6
Comments	6
VBA macros	0
Shapes	2
Dates	1
Cells used	207135
External Links	0
Revision history	0
Tentative assessment	Complex/Dynamic
Percentage formulas	46.2%

Appendix H: STATA Code

```
* Generating Variables
gen cat1 = 0 replace cat1=1 if c3==1|c16==1
gen cat2 =  0 replace cat2=1 if c1==1|c2==1|c13==1|c15==1|c18==1|c19==1
gen cat3 = 0 replace cat3=1 if c4==1|c9==1|c10==1|c14==1
|c17==1|c20==1|c21==1
gen cat4 = 0 replace cat4=1 if c5==1|c6==1|c8==1
gen cat5 = 0 replace cat5=1 if c7==1|c11==1|c12==1

* Descriptives
browse
sum c1 c2 c3 c4 c5 c6 c7 c8 c9 c10 c11 c12 c13 c14 c15 c16 c17 c18 c19
c20 c21

* Exemplary Tabulations
tabulate gender knowledge, row chi2
tab knowledge c1, row chi2
tab knowledge c2, row chi2
tab knowledge c3, row chi2
tab knowledge c4, row chi2
tab knowledge c5, row chi2
tab knowledge c6, row chi2
tab knowledge c7, row chi2
tab knowledge c8, row chi2
tab knowledge c9, row chi2
tab knowledge c10, row chi2
tab knowledge c11, row chi2
tab knowledge c12, row chi2
tab knowledge c13, row chi2
```

```
tab knowledge c14, row chi2
tab knowledge c15, row chi2
tab knowledge c16, row chi2
tab knowledge c17, row chi2
tab knowledge c18, row chi2
tab knowledge c19, row chi2
tab knowledge c20, row chi2
tab knowledge c21, row chi2
tabulate role knowledge, row chi2
tabulate knowledge version, row chi2
tab C21 W, row
tab C21 X, row
tab C21 Y, row
tab C21 Z, row
tab C21 AA, row
tab knowledge cat1, row chi2
tab knowledge cat2, row chi2
tab knowledge cat3, row chi2
tab knowledge cat4, row chi2
tab knowledge cat5, row chi2 sum cat1 cat2 cat3 cat4 cat5
tab knowledge Percentagepercell, row

* Correlation Analysis
corr Percentagepercell knowledge
corr Percentagepercell c8
corr fonts c14
```