Significant Properties of Spreadsheets

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

Remco van Veenendaal
National Archives of the Netherlands

Frederik Holmelund Kjærskov
Danish National Archives

Kati Sein
National Archives of Estonia

Jack O’Sullivan
Preservica

Anders Bo Nielsen
Danish National Archives

Phillip Mike Tømmerholt
Danish National Archives

Jacob Takema
National Archives of the Netherlands

Introduction

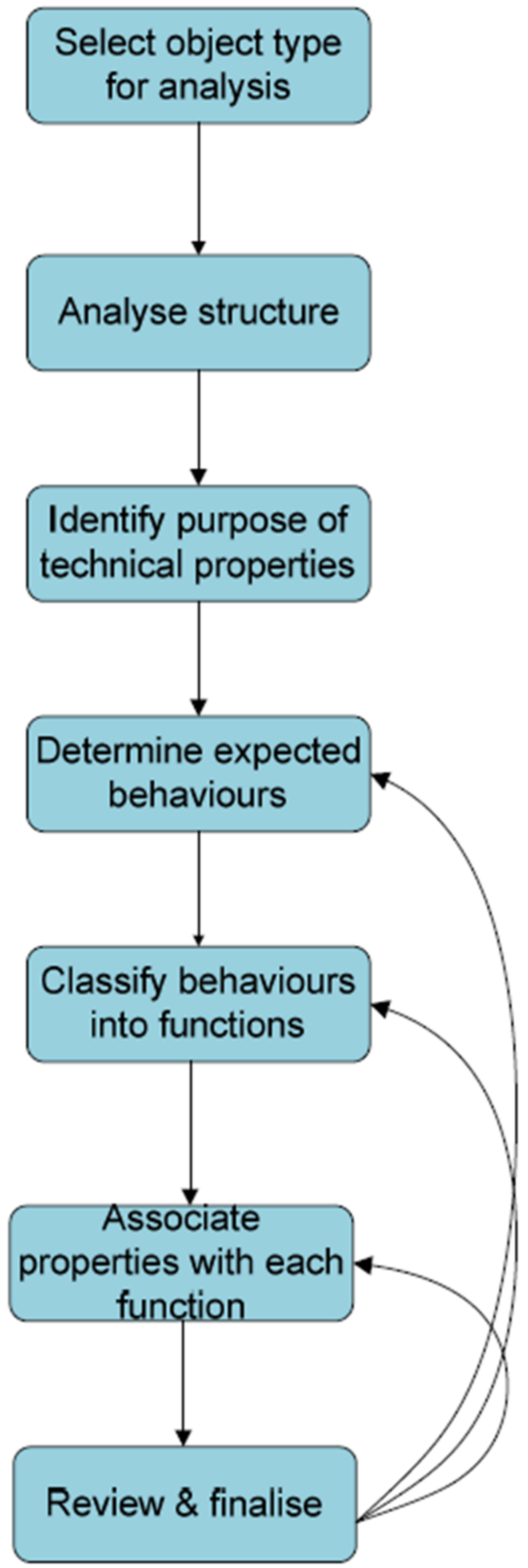
- The Open Preservation Foundation’s Archives Interest Group reports on our ongoing investigation of significant properties of spreadsheets.
- Because we were faced with the **problem of ensuring long-term accessibility of deposited spreadsheets** while preserving their significant properties
- We therefore wanted to **get hands-on experience** in investigating significant properties to understand the original deposited object, and how to preserve it

Method

- We compared methods for identifying significant properties, and chose the methodology for assessing significant properties from the **InSpect Framework Report**.
- It is a formalised, open, standards-based model (based on Function-Behaviour-Structure design method, adapted to this new area of work)
- Used in various (digital preservation) projects, resulting in a Significant Properties Testing Report lore (for Digital Audio Recordings, Structured Text, Raster Images, etc.)

Cooperation

- Monthly calls
- Mailing list
- Google Drive



Select object type for analysis

Analyse structure

Identify purpose of technical properties

Determine expected behaviours

Classify behaviours into functions

Associate properties with each function

Review & finalise

Prerequisites

- Collected a **large set of spreadsheets** from public sources and our collections
- Studied spreadsheet **format specification documents** (VisiCalc, Lotus 1-2-3, Quattro Pro, Microsoft Excel versions, ODS versions, OOXML, Numbers, Gnumeric, …)
- Tested **characterisation tools** (FITS, Apache Tika, DROID, Exiftool, File, JHOVE, NLNZ Metadata Extractor, FIDO, Siegfried, Lingfo, Dependency Discovery Tool, Officeparser.py, Python oletools, …)

Stages

- **Object analysis** (almost done)
  Identify functional, behavioural and structural properties
- **Stakeholder analysis** (to do)
  Analyse functions relevant for a particular user group
- **Reformulation** (where applicable)
  Redevelop object to perform a set of stakeholder functions

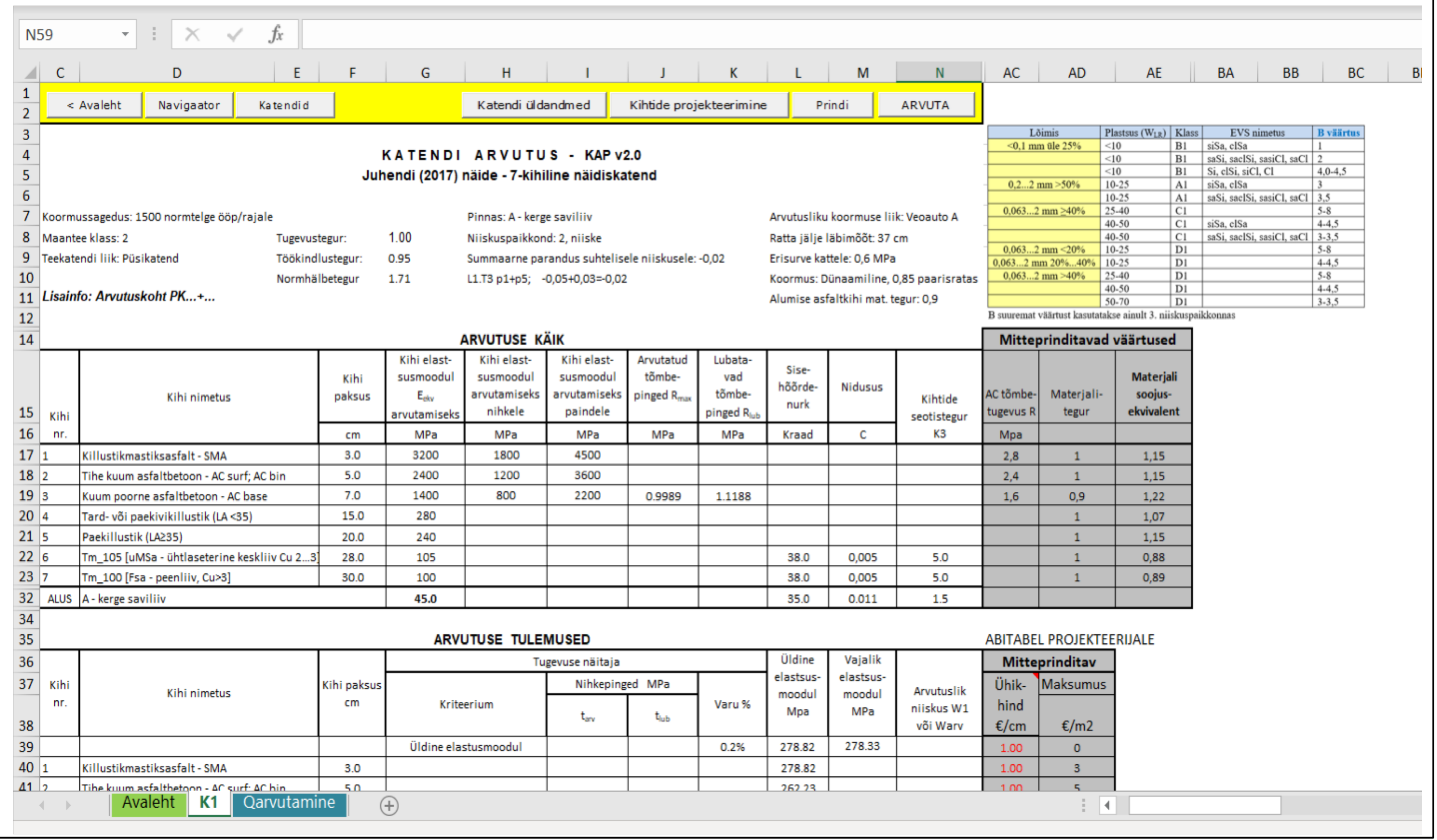
Function, behaviour and structure (properties)

Object type

- Spreadsheets, with 2 tentative subtypes:

1. **‘Simple/static’ spreadsheets** used for (human) visualisation, containing static data values organised into tabular format. Can possibly be migrated to non-spreadsheet file formats.
2. **‘Complex/dynamic’ spreadsheets** contain formulae, notes, macros, dates, links to external data sources or other functions or behaviour. Migrating to non-spreadsheet file formats would cause severe information loss.

A complex/dynamic spreadsheet



KATENDI ARVUTUS - KAP v2.0
Juhendi (2017) näide - 7-kihiline näidiskatend

Koormussagedus: 1500 normtelge ööp/rajale
Maantee klass: 2
Teekatendi liik: Püsikatend
Tugevustegur: 1.00
Töökindlustegur: 0.95
Normhälbetegur 1.71
Pinnas: A - kerge saviliiv
Niiskuspaikkond: 2, niiske
Summaarne parandus suhtelisele niiskusele: -0,02
L1.T3 p1+p5; -0,05+0,03=-0,02
Arvutusliku koormuse liik: Veoauto A
Ratta jälje läbimõõt: 37 cm
Erisurve kattele: 0,6 MPa
Koormus: Dünaamiline, 0,85 paarisratas
Alumise asfaltkihi mat. tegur: 0,9
Lisainfo: Arvutuskoht PK...+...

Lõimis	Plastsus (W_{LR})	Klass	EVS nimetus	B väärtus
<0,1 mm üle 25%	<10	B1	siSa, clSa	1
	<10	B1	saSi, saclSi, sasiCl, saCl	2
	<10	B1	Si, clSi, siCl, Cl	4,0-4,5
0,2...2 mm >50%	10-25	A1	siSa, clSa	3
	10-25	A1	saSi, saclSi, sasiCl, saCl	3,5
0,063...2 mm ≥40%	25-40	C1		5-8
	40-50	C1	siSa, clSa	4-4,5
	40-50	C1	saSi, saclSi, sasiCl, saCl	3-3,5
0,063...2 mm <20%	10-25	D1		5-8
0,063...2 mm 20%...40%	10-25	D1		4-4,5
0,063...2 mm >40%	25-40	D1		5-8
	40-50	D1		4-4,5
	50-70	D1		3-3,5

B suuremat väärtust kasutatakse ainult 3. niiskuspaikkonnas

ARVUTUSE KÄIK

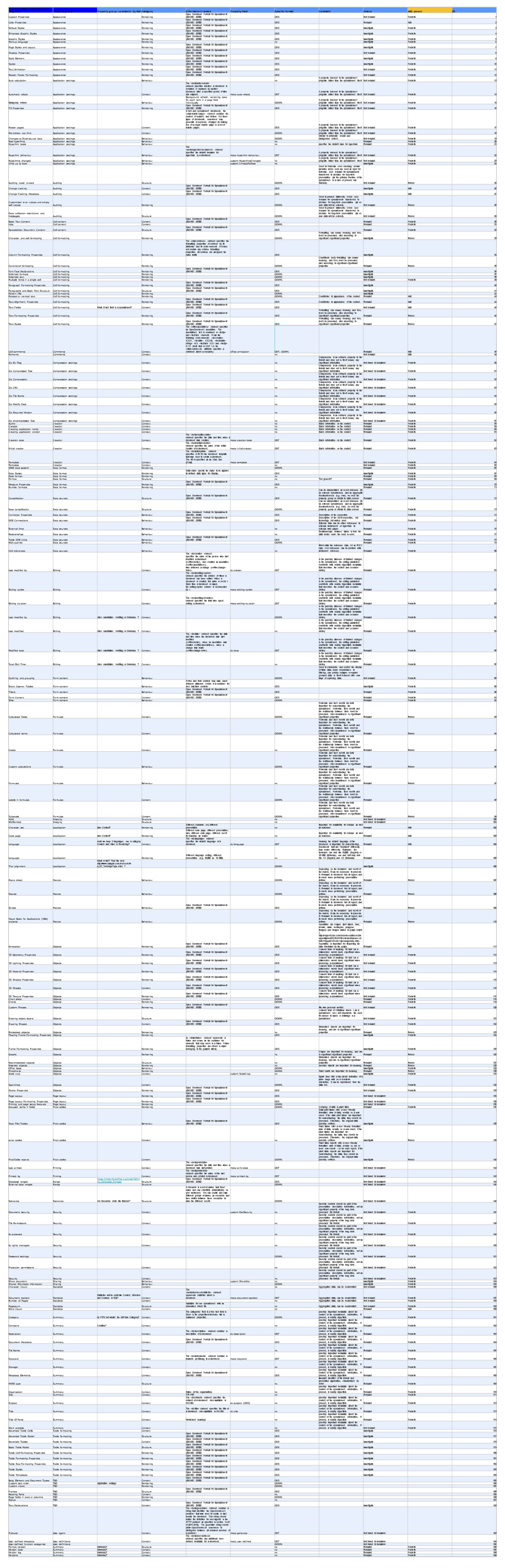
Kihi nr.	Kihi nimetus	Kihi paksus	Kihi elast-susmoodul E_{ekv} arvutamiseks	Kihi elast-susmoodul arvutamiseks nihkele	Kihi elast-susmoodul arvutamiseks paindele	Arvutatud tõmbe-pinged R_{max}	Lubata-vad tõmbe-pinged R_{lub}	Sise-hõõrde-nurk	Nidusus	Kihtide seotistegur K3	Mitteprinditavad väärtused: AC tõmbe-tugevus R	Materjali-tegur	Materjali soojus-ekvivalent
		cm	MPa	MPa	MPa	MPa	MPa	Kraad	C		Mpa		
1	Killustikmastiksasfalt - SMA	3.0	3200	1800	4500						2,8	1	1,15
2	Tihe kuum asfaltbetoon - AC surf; AC bin	5.0	2400	1200	3600						2,4	1	1,15
3	Kuum poorne asfaltbetoon - AC base	7.0	1400	800	2200	0.9989	1.1188				1,6	0,9	1,22
4	Tard- või paekivikillustik (LA <35)	15.0	280									1	1,07
5	Paekillustik (LA≥35)	20.0	240									1	1,15
6	Tm_105 [uMSa - ühtlaseterine keskliiv Cu 2...3]	28.0	105					38.0	0,005	5.0		1	0,88
7	Tm_100 [Fsa - peenliiv, Cu>3]	30.0	100					38.0	0,005	5.0		1	0,89
ALUS	A - kerge saviliiv		**45.0**					35.0	0.011	1.5			

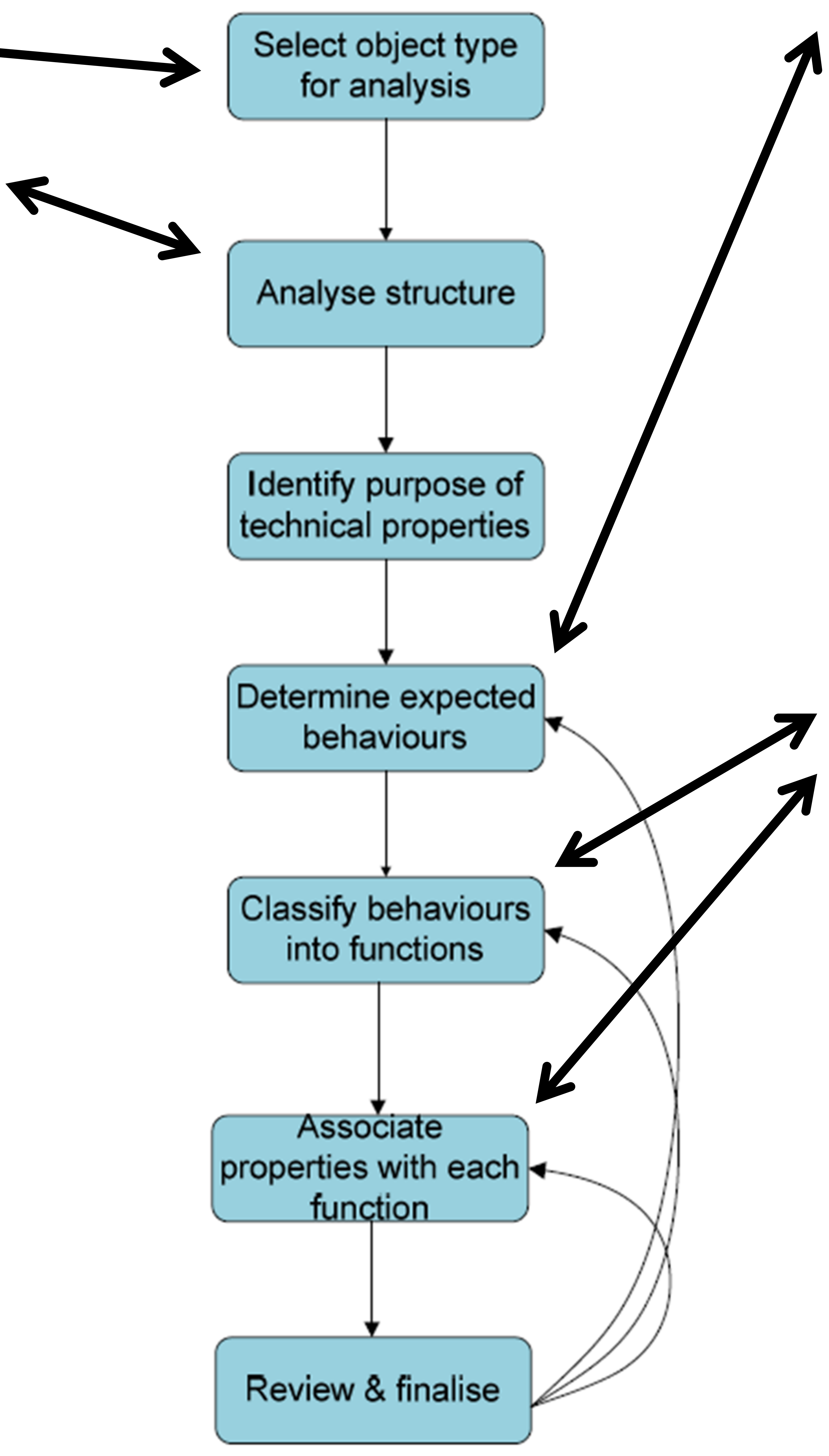
ARVUTUSE TULEMUSED

ABITABEL PROJEKTEERIJALE

Kihi nr.	Kihi nimetus	Kihi paksus cm	Tugevuse näitaja: Kriteerium	Nihkepinged MPa: t_{arv}	t_{lub}	Varu %	Üldine elastsus-moodul Mpa	Vajalik elastsus-moodul MPa	Arvutuslik niiskus W1 või Warv	Mitteprinditav: Ühik-hind €/cm	Maksumus €/m2
			Üldine elastusmoodul			0.2%	278.82	278.33		1.00	0
1	Killustikmastiksasfalt - SMA	3.0					278.82			1.00	3
2	Tihe kuum asfaltbetoon - AC surf; AC bin	5.0					262.23			1.00	5

Properties





Select object type for analysis

Analyse structure

Identify purpose of technical properties

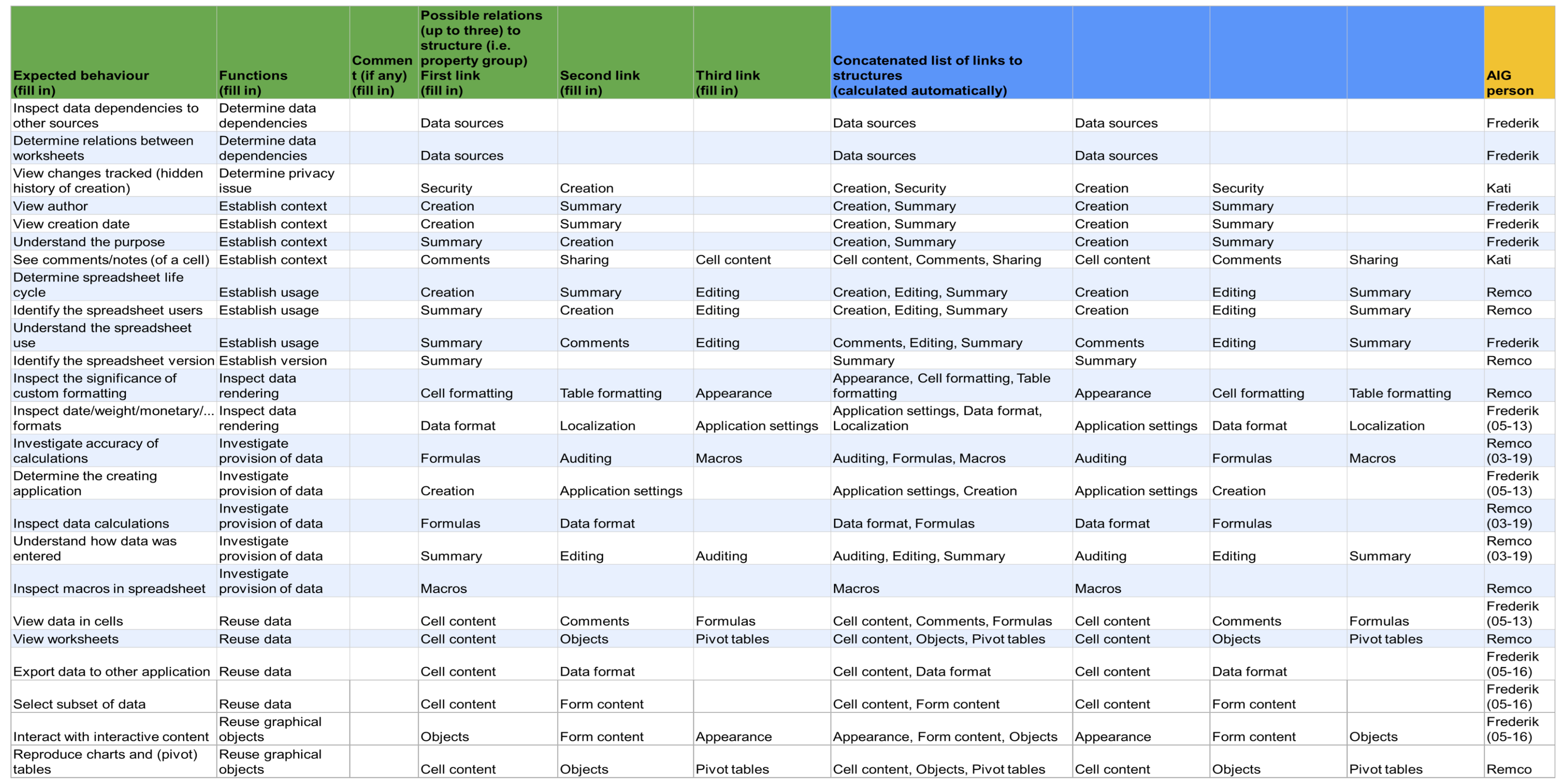
Determine expected behaviours

Classify behaviours into functions

Associate properties with each function

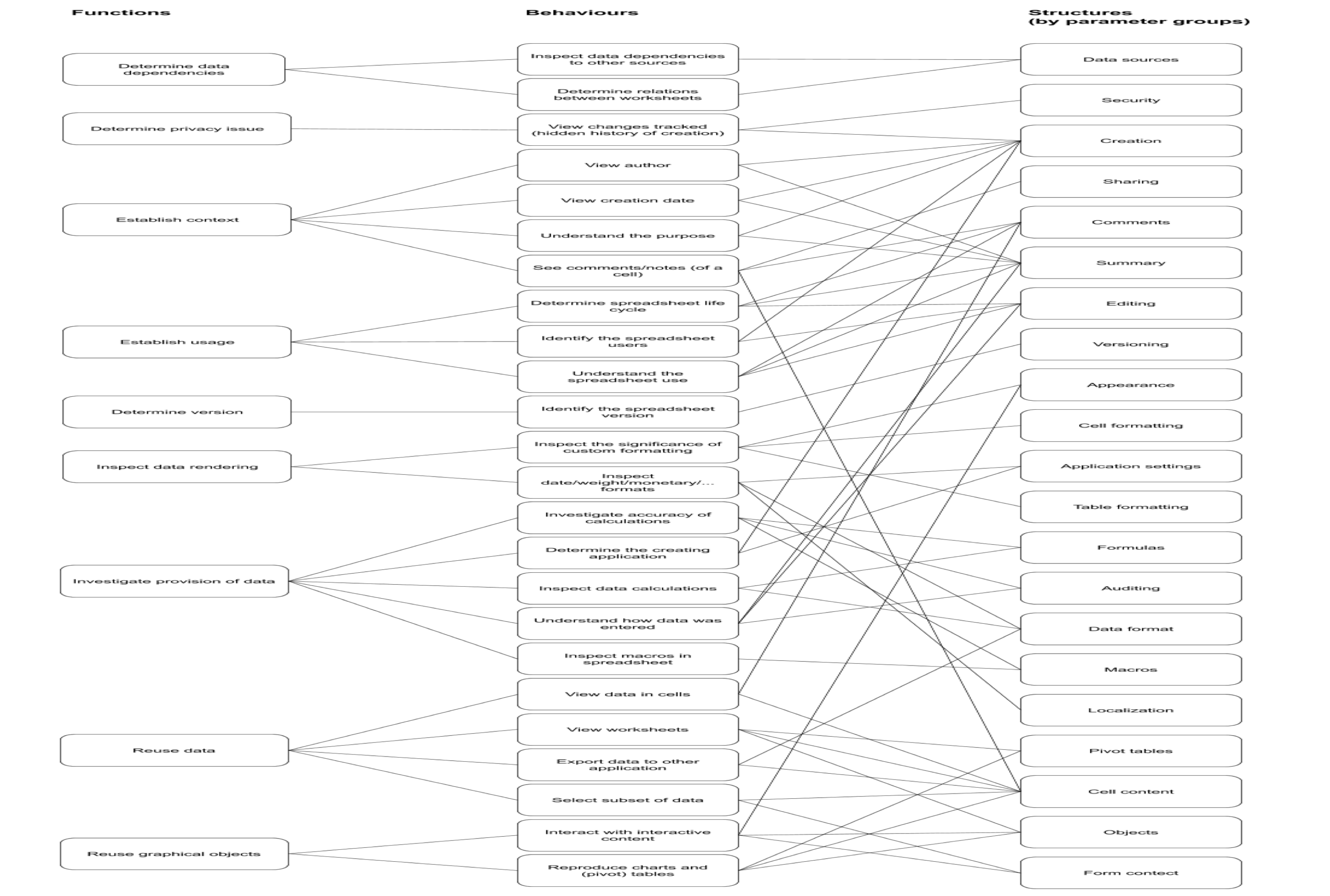
Review & finalise

Expected behaviours



Expected behaviour (fill in)	Functions (fill in)	Comment (if any) (fill in)	Possible relations (up to three) to structure (i.e. property group) First link (fill in)	Second link (fill in)	Third link (fill in)	Concatenated list of links to structures (calculated automatically)				AIG person
Inspect data dependencies to other sources	Determine data dependencies		Data sources			Data sources	Data sources			Frederik
Determine relations between worksheets	Determine data dependencies		Data sources			Data sources	Data sources			Frederik
View changes tracked (hidden history of creation)	Determine privacy issue		Security	Creation		Creation, Security	Creation	Security		Kati
View author	Establish context		Creation	Summary		Creation, Summary	Creation	Summary		Frederik
View creation date	Establish context		Creation	Summary		Creation, Summary	Creation	Summary		Frederik
Understand the purpose	Establish context		Summary	Creation		Creation, Summary	Creation	Summary		Frederik
See comments/notes (of a cell)	Establish context		Comments	Sharing	Cell content	Cell content, Comments, Sharing	Cell content	Comments	Sharing	Kati
Determine spreadsheet life cycle	Establish usage		Creation	Summary	Editing	Creation, Editing, Summary	Creation	Editing	Summary	Remco
Identify the spreadsheet users	Establish usage		Summary	Creation	Editing	Creation, Editing, Summary	Creation	Editing	Summary	Remco
Understand the spreadsheet use	Establish usage		Summary	Comments	Editing	Comments, Editing, Summary	Comments	Editing	Summary	Frederik
Identify the spreadsheet version	Establish version		Summary			Summary	Summary			Remco
Inspect the significance of custom formatting	Inspect data rendering		Cell formatting	Table formatting	Appearance	Appearance, Cell formatting, Table formatting	Appearance	Cell formatting	Table formatting	Remco
Inspect date/weight/monetary/... formats	Inspect data rendering		Data format	Localization	Application settings	Application settings, Data format, Localization	Application settings	Data format	Localization	Frederik (05-13)
Investigate accuracy of calculations	Investigate provision of data		Formulas	Auditing	Macros	Auditing, Formulas, Macros	Auditing	Formulas	Macros	Remco (03-19)
Determine the creating application	Investigate provision of data		Creation	Application settings		Application settings, Creation	Application settings	Creation		Frederik (05-13)
Inspect data calculations	Investigate provision of data		Formulas	Data format		Data format, Formulas	Data format	Formulas		Remco (03-19)
Understand how data was entered	Investigate provision of data		Summary	Editing	Auditing	Auditing, Editing, Summary	Auditing	Editing	Summary	Remco (03-19)
Inspect macros in spreadsheet	Investigate provision of data		Macros			Macros	Macros			Remco
View data in cells	Reuse data		Cell content	Comments	Formulas	Cell content, Comments, Formulas	Cell content	Comments	Formulas	Frederik (05-13)
View worksheets	Reuse data		Cell content	Objects	Pivot tables	Cell content, Objects, Pivot tables	Cell content	Objects	Pivot tables	Remco
Export data to other application	Reuse data		Cell content	Data format		Cell content, Data format	Cell content	Data format		Frederik (05-16)
Select subset of data	Reuse data		Cell content	Form content		Cell content, Form content	Cell content	Form content		Frederik (05-16)
Interact with interactive content	Reuse graphical objects		Objects	Form content	Appearance	Appearance, Form content, Objects	Appearance	Form content	Objects	Frederik (05-16)
Reproduce charts and (pivot) tables	Reuse graphical objects		Cell content	Objects	Pivot tables	Cell content, Objects, Pivot tables	Cell content	Objects	Pivot tables	Remco

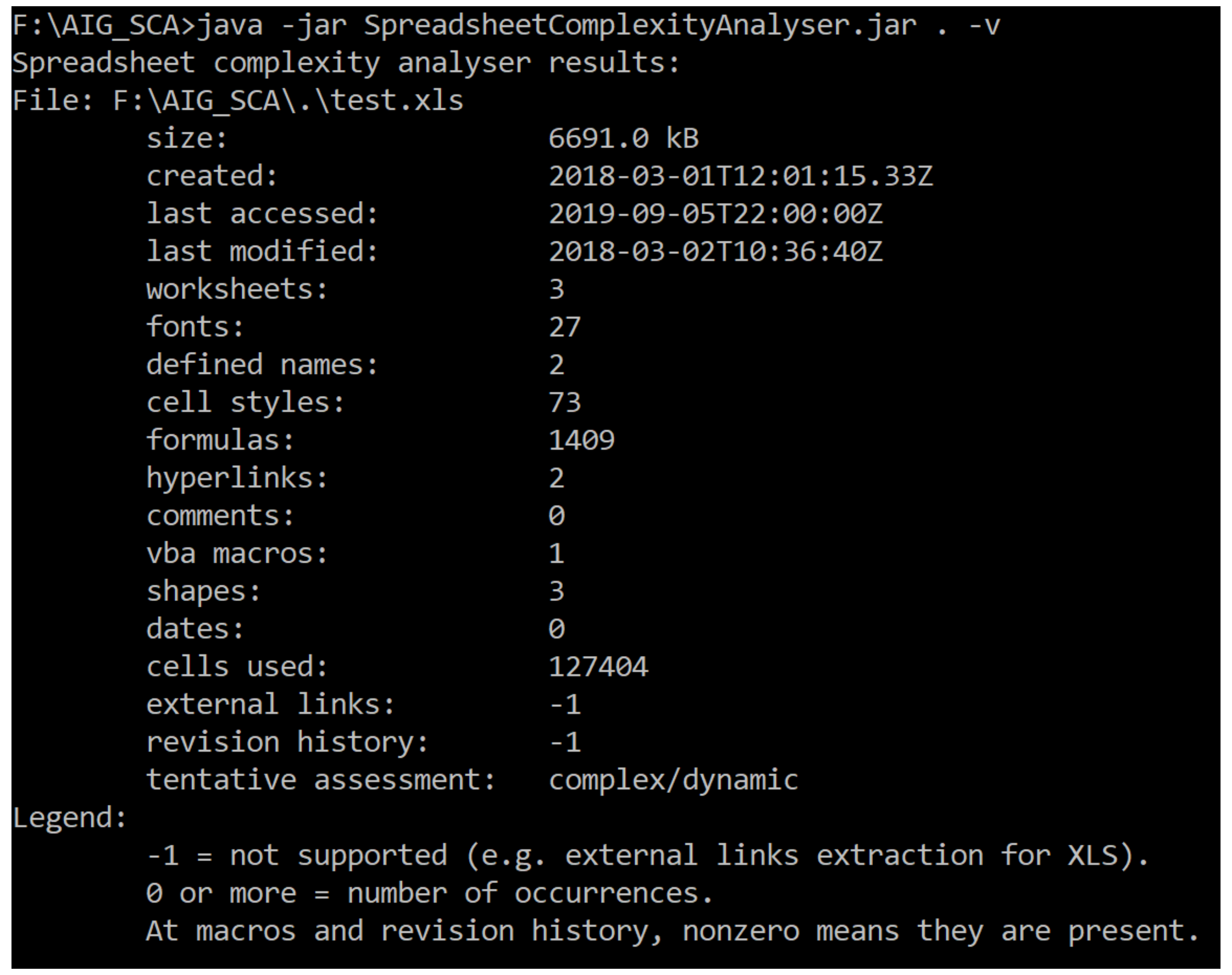
FBS diagram


Functions
Behaviours
Structures
(by parameter 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

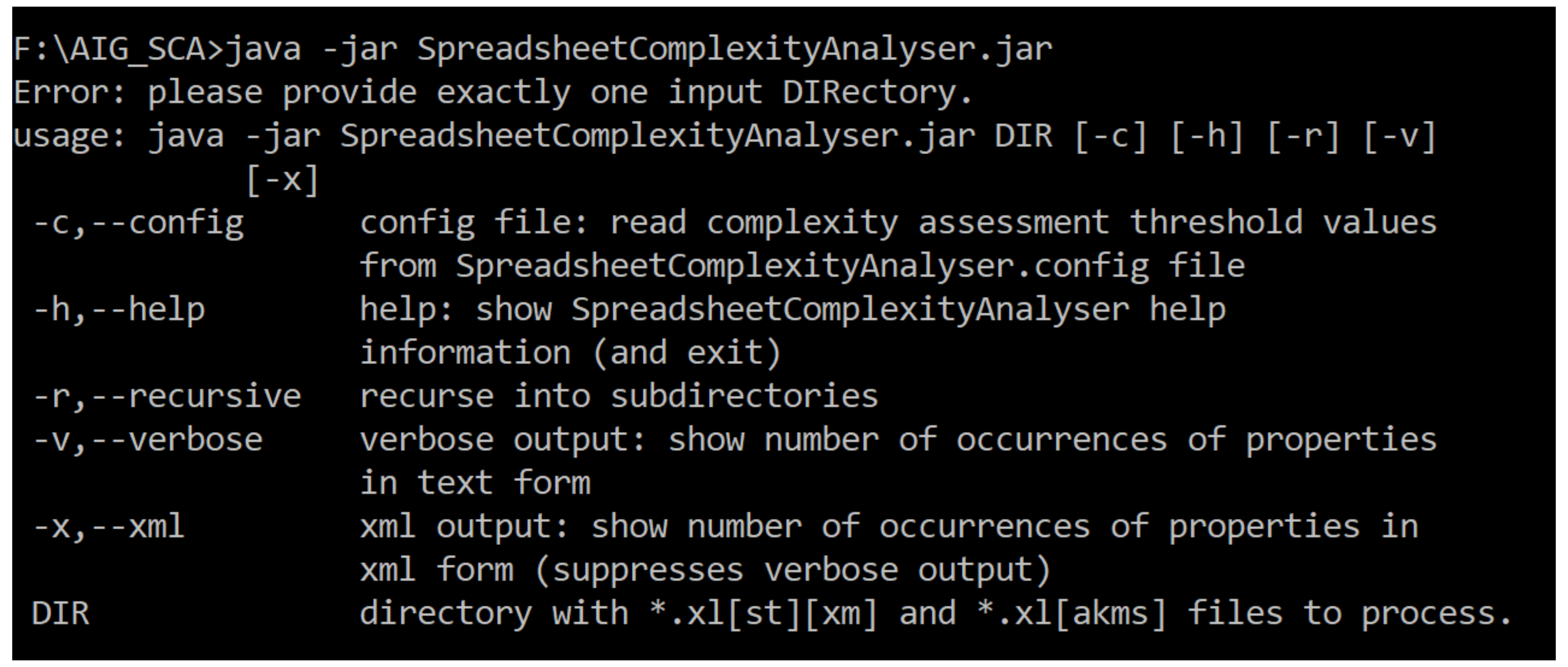
Spreadsheet Complexity Analyser and conclusion

Spreadsheet Complexity Analyser

- Prototype (CC0) **open source tool**
- Extracts **spreadsheet-specific properties**
  Workbook: worksheets, fonts, defined names, cell styles, external links and revision history
  Sheet (totaled up): formulas, hyperlinks, comments, shapes, dates, cells used
  VBA: nonzero indicates possible vba macros
  File: file size, creation date/time, last accessed, last modified
- Assesses 'complexity' (using default or user thresholds)



```
F:\AIG_SCA>java -jar SpreadsheetComplexityAnalyser.jar . -v
Spreadsheet complexity analyser results:
File: F:\AIG_SCA\.\test.xls
        size:                   6691.0 kB
        created:                2018-03-01T12:01:15.33Z
        last accessed:          2019-09-05T22:00:00Z
        last modified:          2018-03-02T10:36:40Z
        worksheets:             3
        fonts:                  27
        defined names:          2
        cell styles:            73
        formulas:               1409
        hyperlinks:             2
        comments:               0
        vba macros:             1
        shapes:                 3
        dates:                  0
        cells used:             127404
        external links:         -1
        revision history:       -1
        tentative assessment:   complex/dynamic
Legend:
        -1 = not supported (e.g. external links extraction for XLS).
        0 or more = number of occurrences.
        At macros and revision history, nonzero means they are present.
```



```
F:\AIG_SCA>java -jar SpreadsheetComplexityAnalyser.jar
Error: please provide exactly one input DIRectory.
usage: java -jar SpreadsheetComplexityAnalyser.jar DIR [-c] [-h] [-r] [-v]
            [-x]
 -c,--config      config file: read complexity assessment threshold values
                  from SpreadsheetComplexityAnalyser.config file
 -h,--help        help: show SpreadsheetComplexityAnalyser help
                  information (and exit)
 -r,--recursive   recurse into subdirectories
 -v,--verbose     verbose output: show number of occurrences of properties
                  in text form
 -x,--xml         xml output: show number of occurrences of properties in
                  xml form (suppresses verbose output)
 DIR              directory with *.xl[st][xm] and *.xl[akms] files to process.
```

Acknowledgements

- Thank you to Becky McGuinness, Charlotte Armstrong and Carl Wilson of OPF for supporting our work

Conclusion

- OPF AIG is using the InSpect methodology to investigate the significant properties of spreadsheets. We have almost finished the Object analysis stage.
- A Spreadsheet Testing Report will be added to the InSpect Testing Report lore.
- The Spreadsheet Complexity Analyser tool extracts spreadsheet properties and assesses complexity.

- Preliminary conclusions support earlier findings of significant property studies:
- The complexity of and degree of freedom inherent in spreadsheets makes creating an **exhaustive list of significant spreadsheet properties practically impossible**. But a list of (technical) significant properties does help choose suitable file formats
- To obtain such lists, one needs to **identify relevant stakeholders**, i.e. complete stakeholder analysis

References

1. Bechhofer, S., Sierman, B., Jones, C., Elstrøm, G., Kulovits, H., Becker, C.: Final version of policy specification model. http://www.scape-project.eu/deliverable/d13-2-catalogue-of-preservation-policy-elements (2014). Accessed March 15, 2019
2. Knight, G.: InSPECT Framework Report. https://web.archive.org/web/20160520083956/http://www.significantproperties.org.uk/inspect-framework.html (2009). Accessed March 15, 2019
3. Lucker, P., Sijtsma, C., van Veenendaal, R.: Significant Significant Properties. https://osf.io/rtjw3 (2018). Accessed March 15, 2019