JACOBS
UNIVERSITY

Andrea Kohlhase

Framings of Information: Readers' Perception of Information Sources in Spreadsheets

Technical Report No. 30
May 2013

School of Engineering and Science

Framings of Information

Readers' Perception of Information Sources in Spreadsheets

Andrea Kohlhase

School of Engineering and Science
Jacobs University Bremen gGmbH
Campus Ring 12
28759 Bremen
Germany

E-Mail: a.kohlhase@jacobs-university.de
http://www.jacobs-university.de/

Summary

Spreadsheets have become very popular tools for analyzing and visualizing data from business and science. Unfortunately, error rates and and misinterpretations have increased dramatically with their complexity and impact over the years. To better understand human-spreadsheet interaction, in this paper we explore readers' information models, but in contrast to most studies we focus on spreadsheet readers rather than spreadsheet authors. In particular, we investigate the perception of information sources in spreadsheets. We conducted 14 repertory grid interviews and analyzed them with the help of a Generalized Procrustes Analysis. The results suggest several information framings in spreadsheets from the point of view of spreadsheet readers: distinct information- and interface perception, information quality varying between data and knowledge, role-specificity of information objects, information dependency on neighboorhood, and an (unexpected) document/player metaphor for spreadsheets. Finally, we envision new human-spreadsheet interactions to increase the readibility, thus usability of spreadsheets.

Contents

1 Introduction

The intuitive, flexible, and direct approach to computation in spreadsheets has led to widespread use and reuse. In particular, spreadsheet programs[1] have become very popular to create, modify, and visualize numeric business and science data. In turn complexity and impact increased dramatically over the years. It has been estimated that each year tens of millions professionals and managers create hundreds of millions of spreadsheet programs [24]. This intensity yields not only more and more shared, complex spreadsheet programs, but also wide-impact errors on the data level (up to 90% [24], see also [27]) and on the comprehension level (e.g. [26]). The losses caused by formula errors and misinterpretation have even led to an international task force to battle them [5]. Therefore, **human-spreadsheet interaction**, i.e., the process and objects of interaction of humans with spreadsheet programs, is an emerging field of investigation.

1.1 Approach: Focusing on Spreadsheet Readers

Based on cognitive psychology Lewis and Olson gave a critical account for the success of spreadsheets [17] in terms of users. In particular, the barriers to programming are lowered, since the spreadsheet model can be used as visual programming language, enabling programming with low entry costs and early experience of success through effective displays and operations. Therefore, spreadsheets became the paradigmatic example for **end-user programming (EUP)**, i.e., "programming to achieve the result of a program primarily for personal [...] use" [15, p. 4][2], and an abundance of research in EUP draws on this. From the point of view of EUP, a spreadsheet user is an end-user, that is "simply any computer user" [ibd] who creates a spreadsheet program.

But Nardi and Miller noticed another feature of spreadsheets in [21]: they are not "single-user applications". In particular, they are used in the work environment as a communication and collaboration tool to exchange or combine domain knowledge and programming expertise. Even though Nardi and Miller conclude that spreadsheets are used for communication, they still only look at the collaborative aspect of this communication, that is, the creation of spreadsheets. Therefore here, spreadsheet users are classified according to their different programming skills into non-programmers, local developers, and programmers [ibd,

[1] In this paper we distinguish between "**spreadsheet program**" and "**spreadsheet application**". The former refers to an instance of the media type of spreadsheets, whereas the latter points to the environment handling spreadsheet programs. Note that a spreadsheet program is no ordinary software program as the difference between editing and executing is veiled by a full integration with the spreadsheet application, that in turn becomes a development as well as a presentation environment. Sometimes, when we do not want to distinguish between the two, we use "spreadsheet(s)" to encompass both – as is common in everyday language.

[2] This definition differs from earlier ones involving low expertise and experience programming levels, which is of no consequence here. We refer the interested reader to a discussion in [15].

p. 201] and kept up in following research.

In [12] Hendry and Green were able to verify weaknesses of spreadsheets in use (already analyzed in [17]). They highlighted the fact that spreadsheet use is also a matter of understanding. In particular, they report their informants' missing comprehension (even of their own spreadsheet) and trace it to lacking comprehensibility support by spreadsheets. For example, secondary notation – the presentation of extra information via notation conventions – is carried "primarily by choice of layout" [12, p. 1065]. Spreadsheets are also discussed as information devices. In their user study though, the informants were asked "to explain how their own spreadsheets worked" [12, p. 1033]. This indicates, that here as well spreadsheet users are perceived as authors of spreadsheets.

The standard research on usability problems of spreadsheets is based on Panko's influential report on error states and types in [25]. He also only considers errors that spreadsheet authors introduced, e.g. computational errors based on faulty formulae.

We can summarize that the main research strains with respect to spreadsheets are restricted by their concern for spreadsheet authors only. Notable exceptions are [11, 16, 12, 31]. But are these really the only users of spreadsheets (see e.g. [1] for an overview of spreadsheet users and use)? What about people who

- make use of existing templates by simply putting in new data?
- review data developments on different abstraction levels e.g. supervisors or members of a board?
- assess data to base further decisions upon (see [2])?
- want to understand their own spreadsheet program after a while?
- look for reusable parts of a spreadsheet program, therefore browsing available ones?

They are spreadsheet users, but not spreadsheet authors. We call them "**spreadsheet readers**" or for short "readers" to stress this fact. Several studies indicate that not only computation, but also presentation of data is a major aim of spreadsheet use, see e.g. [3, 2, 29, 12, 22].

In a nutshell, we can say that research with a focus on spreadsheet readers is necessary, but still largely missing. Therefore, in this paper, we focus on spreadsheet readers and investigate, how people perceive spreadsheets when reading them.

1.2 Contribution: Readers' Information Model of Spreadsheets

From the point of view of spreadsheet readers, at first glance, spreadsheets can be considered *data interfaces* that display and allow to play with data. But the many reports about bad decisions caused by misinterpretation and difficulties of spreadsheet comprehension allow sensible doubts about the completeness of this approach. Data by itself is not interesting, only if it becomes information or even

knowledge it is usable.

In this paper, we investigate what representations are used in spreadsheets to turn data into something more valuable. Since the evaluation of any representation "depends on what you want to use it for" [7, p. 6], we cannot turn to Green and Petre's cognitive dimensions framework for visual programming environments [ibd], as it primarily works as a discussion tool of software artifacts for programmers.

The spreadsheet reader perceives such representations as distinguishable interface objects that carry information ("**information objects**"), which together with their relations ("**information framings**") built up an **information model**. Therefore, on the one hand we study the available set of information objects in the spreadsheet interface, on the other we investigate their relations – both from the point of view of spreadsheet readers but with a focus on the latter.

The contribution of this paper consists of a turn towards the spreadsheet reader in terms of spreadsheet research, a first identification and classification of information objects in spreadsheets from this point of view, and mainly an exploration of the readers' information model. Based on the latter we suggest new interactions to increase the usability of spreadsheets.

1.3 Methodology: Repertory Grids and General Procrustes Analysis

To better understand what spreadsheet readers perceive as information units, what meaning they assign to these information objects, and how they discriminate between them, we conducted a study using the **Repertory Grid Interview (RGI) Technique** [14, 13].

RGI explores personal constructs, i.e., how persons perceive and understand the world around them. McKnight was the first to suggest RGI for exploration of an information space [18], Newby suggested a statistical method based on eigenvalue construction to align cognitive space and information space [23], turning RGI into a semi-empirical method. By now RGI is a well-established method to explore users' personal constructs when interacting with software artifacts (see [30] for a list of examples). One advantage over other methods is that a "small sample size is commonly used when implementing a repertory grid investigation [... e.g. for] a given population, the use of ten participants will ensure determination of the complete set of important constructs" [4].

A **repertory grid** is a grid consisting of "**elements**", i.e., the objects under consideration, and "**constructs**", i.e., pairs of antithetical properties that separate elements. The constructs serve as a bipolar dimension on which the elements are evaluated. As the property elicited first in a construct is the more salient one, RGI calls it the "**implicit pole**" and the other one emerging in the reflection of the dimension of comparison the "**emergent pole**". Elements as well as constructs can be elicited from the test persons themselves or can be provided by the

interviewer. Comparison of multiple repertory grids is simplified if the individual ratings are given on a fixed set of elements or/and constructs, but a free elicitation explores the cognitive space.

For our main RGI we decided to fix the set of elements, but to elicit individual constructs to better understand the information space. Concretely, we conducted a pilot study to determine half of the element set, consisting of information objects commonly recognized by spreadsheet readers. Then we added a set of additional, non-standard spreadsheet information units to potentially broaden the boundary of the spreadsheet information space. Specifically, we added constructs suggested by a semantic help system for `MS Excel` '03 programs called "`SX`" [16]. We were specifically interested to find out which of these external information objects were perceived to deliver similar or different information compared to traditional spreadsheet information objects.

We analyzed the repertory grid data obtained in the main study with "Idiogrid" [9]. We performed Gower's **Generalized Procrustes Analysis (GPA)** [6], as it can be used when data "have arisen from one type of scaling of the same stimuli as perceived by different individuals" [ibid, p. 33]. In particular, we followed the analysis as described by Grice in [8]. With GPA, three-dimensional data matrices can be analyzed with a multivariate statistical technique. In particular, in our RGI we can compare the individual (dimension 1) natural language constructs (dimension 2) rated on our fixed set of information objects (dimension 3).

In GPA the first step is the construction of an average grid from all rating grids after an approximate alignment via Procrustes rotation, yielding the "**consensus grid**", i.e., a best fit grid for a number of grids that are equal in one dimension but not in the other. A randomization test and subsequent standard ANOVA analysis gives us e.g. information about its statistical significance. Note that this randomization test ensures the validity of the consensus proportion against the Procrustes rotation sensitivity especially with small data sets (see a discussion in [10]).

A subsequent Principal Components Analysis (PCA) analyzes the correlations of the consensus grid and the components score coefficients are saved for later use. Then a concatenated grid is built from the consensus grid and all individual repertory grids and exposed to an extension analysis. Here, the components that were created earlier in the PCA are re-constructed and the individual constructs are mapped into the resp. PC_iPC_j biplot space. This way we obtain a visual presentation of the most salient constructs with respect to the principal components PC_i and PC_j. With a qualitative analysis these can then be interpreted as distinct framings of information resulting in an exploration of readers' information model.

First, we present the setup of our repertory grid interviews, then we present and interpret results towards a cognitive dimension framework of spreadsheet readers. Finally, we suggest spreadsheet programs to advance from data interfaces to knowledge interfaces to overcome usability issues especially for spreadsheet read-

ers.

2 The RGI Study

The aim of the study is a better understanding of (existing and potential) information conveyed with a spreadsheet, or in other words, to explore the spreadsheets' information space as seen by spreadsheet readers.

2.1 The Pilot Study: A Fixed Set of Common Information Objects in Spreadsheets

In a first RGI we explored which information objects were discerned by spreadsheet readers in common spreadsheets. From this we extracted the most relevant ones to be included in the fixed set of elements for our main RGI study.

In the pilot study 4 subjects (bachelor student, master student, PhD student, professor of computer science) participated. Neither had written spreadsheet programs professionally, i.e., in a more sophisticated way than making use of standard functions in formulae and of standard spreadsheet functionality. Therefore, we consider them as typical (technically oriented) spreadsheet readers.

We presented each subject a simple, but complexly structured spreadsheet (on a laptop). The interviewer then asked the participant to nominate information objects, that is, objects that carry information, in this particular spreadsheet. Each labelled information object was explained to the interviewer, written onto a paper board card ("element card") by the subject, and put into a (paper) grid as column headers by the interviewer. Following traditional RGI, the interviewee was then handed three randomly selected element cards and asked to name one way in which two of the selected elements – considered as information objects – are similar or different from the other one. The label for the sameness was noted in the grid as left row header (the emergent pole), the label for the difference as right row header (the implicit pole) - yielding a construct. Then all elements were evaluated with respect to this construct with a binary rating scale: does this element rather belong to the implicit pole or the emergent pole? All in all, 43 elements and 43 constructs were collected, each interviewee contributing from 9-11 and 10-11 respectively.

Thinking about the constructs often triggered a reidentification of the elements, some were discarded, others added or renamed. The sessions took between three and four hours and were exhausting for the participants. In the end each interviewee had the feeling that the elicited elements fully described 'the' information space offered by spreadsheets. We were surprised by the richness and individuality of perceptions. In order to extract the most dominant information objects for spreadsheet readers out of the 43 given element labels, we categorized them via

Table 1: Common Information Objects of Spreadsheets

`Title`	A phrase describing the content of the spreadsheet
`Headers`	A (short) phrase supporting the interpretation of values of a regionally close range of cells (e.g. a column header)
`Legends`	A list of content properties and resp. layouts (as in a map legend)
`Values`	The content of a cell container
`Formulae`	A computational rule that yields a cell value
`(sx:)Color Coding`	The use of color hinting at additional information
`Tables`	A possibly multidimensional homogenous structural layout of cells, that is perceived as an object of its own

the used frame — which is a tremendous oversimplification for each subject's individual information space, but which was required for a manageable main study with a focus on constructs. For example, let us look at the frame of "grid-like substructures in a worksheet": all participants listed elements according to this frame, but they referred to "Block", "Row", "Column", "Table", or "Independent Subtable". When several of these were listed by one subject, the evaluation wrt. the constructs turned out to be very similar, so we felt justified to join them. We found six information frames to be consistently listed (see Table 1). They were embodied by labels to gain specific information objects. Note that "diagrams" are missing, which may be due to the fact, that they were not part of our standard spreadsheet example. Subsequently, we asked our participants to assess the mappings between their personal information space and the one represented by these 7 elements. They complained about its missing sophistication, but confirmed it (with a heavy heart).

2.2 A Fixed Set of Complementary Information Objects in Spreadsheets

To broaden this set of elements, we added information objects which are not traditionally used in spreadsheets. In particular, we looked for such objects that contain spreadsheet-related information not usually available to spreadsheet readers. For this we made use of the spreadsheet extension "`SX`" [16].

`SX` aims at providing user assistance for spreadsheet readers based on a background ontology. As "cells" are important information objects in spreadsheets, `SX` acts cell-oriented. In a big spreadsheet a reader clicks for example on a cell that contains "444" as information of `Values`. Common information objects tell her about the context (e.g. by `Title` "Loss and Profit Statement", `Headers` "Profits" and "2011", or `Legends` "in Millions"). But what if she doesn't understand how "Profits" are calculated? When using `SX` this cell might be linked to a concept

Table 2: Extra Information Objects of Spreadsheets

sx:Localized Info	A local look-up (data and text) of relevant information for cells on a by-cell-click basis
sx:Functional Block	A local border indicating all cells functionally associated to the currently selected cell
sx:Dependency Graph	An overview graph (in a different window) of concepts showing on which the corresponding (selected) cell is ontologically dependent
sx:Relational Arrows	An arrow indicating a dependency relation between concepts in sx:Dependency Graph
sx:Concept Nodes	A node in sx:Dependency Graph representing a dependent subconcept, that additionally serves as a link to corresponding spreadsheet cells

in a background ontology that covers the domain knowledge of this particular spreadsheet. The reader likes to retrieve this linked concept from the ontology by opting for a "look-up" option provided by SX and by selecting the wanted cell. A pop-up close to the selected cell will appear with this additional information — e.g. "A profit is the difference between revenues and expenses." — together with the header information "Profits [2011]".

A group of SX experts identified the information objects in Table 2 as most relevant and dissimilar to common information objects in spreadsheets SX resources. They added (sx:)Color Coding as in Table 1 for relevance. The union of both sets of information objects were used as the given, fixed set of elements, for which in our second RGI study constructs were to be elicited.

2.3 The Main RGI Study: Participants and Procedures

For our investigation we interviewed 14 people, of which 10 were male and 4 female. The age distribution was the following:

Age	≤ 20	≤ 30	≤ 40	≤ 50
	5	6	2	1

One subject had authored spreadsheets on a professional basis, 4 subjects were familiar with authoring simple spreadsheets, the other 9 only had occasional contact. All were explicitly asked to take up the role of a spreadsheet reader. Their background and education varied, but 3 were familiar with the MS Excel add-in SX before the interview.

The procedure for the elicitation of the constructs was the same as described for the pilot study in Section 2.1 except for an introduction of the fixed set of information objects. Each element was explained by the interviewer and SX was introduced where necessary.

The rating scale was essentially binary: it consisted of -1,0,1 but the interviewees were only told about their option to use "0" as a rating when they otherwise would have discarded the construct in question as inapplicable. In 1,5 to 3hr sessions participants reported an average of 8.2 construct pairs (SD = 1.4) ranging between 5 and 11 pairs. A total of 115 constructs were elicited.

We focused each repertory grid by swapping the construct poles to optimize the amount of applicable poles for the set of common spreadsheet information objects. This way we could identify the characteristic construct poles for common versus complementary elements and the according pole distribution.

3 Statistical Results and Discussion

We now give an overview of the statistical analysis of the RGI study data via a General Procrustes Analysis executed in Idiogrid. This is followed by an interpretative discussion of the findings.

Figure 1: Extension Analysis Biplot for PC_1 and PC_2 (with Element Clusters)

The first component of a GPA is the computation of the consensus grid of all individual repertory grids. The consensus proportion turned out to be .68, which indicates a rather high similarity. It was tested for statistical significance with the help of a randomization test based on 500 trials, which yielded an observed proportion $p \leq 0.00$. This verifies that the consensus grid contains statistically significant data, that are worthwhile to be analyzed further.

The second step of a GPA consists in a standard Principal Components Analysis (PCA) on the consensus grid yielding components $\{PC_{i=1,\ldots,11}\}$. The first component explains ca 33.7%, the second 22.2% and the third 14.4% of the variance in the data.

In Fig. 1 we can see the outcome of the extension analysis (containing the reconstructed first three PCs of the consensus grid and mapped elements and individual constructs) for PC_1 and PC_2 (with only the more salient constructs, in particular with 0.84% suppression of labels) run by Idiogrid. Emergent Poles are marked by a "(-)" prefix. A simple concatenation of all individual repertory grids was exposed to a cluster analysis in OpenRepGrid[3] yielding Fig. 2. The distinguished clusters are displayed as fenced regions in Fig. 1.

Figure 2: Element Cluster Dendrogram for the Concatenated Grid

To approximate the meaning of the principal components, we looked at elicited similar constructs, that is at the more salient ones close to the axes in Fig. 1. Then we tried to find categories that can serve as common denominator constructs. As this content analysis was qualitative, the reliability was ensured by following the procedure given in [13, 155ff.].

[3] `http://www.openrepgrid.uni-bremen.de/wiki`

3.0.1 Principal Component PC_1

Probst et al. suggested in [28] a knowledge management model positing that GLYPHS, DATA, INFORMATION, and last but not least KNOWLEDGE can be seen as stages of a pipeline as in Fig. 3. This model differentiates what we have simply

Figure 3: Knowledge Management Model after [28]

called "information" into four distinct traits. GLYPHS are just a set of characters without any structure, combined with a syntax they become DATA, additionally enriched by context they become INFORMATION, and finally, they turn into KNOWLEDGE if a semantic net or a global context is present.

The constructs coming closest to the first principal component (depicted by the horizontal axis in Fig. 1) are:

Implicit Pole	**Emergent Pole**
meta level	(-)object level
relevant for analysis	(-)relevant for understanding
dependency info	(-)not formal info
outside of spreadsheet	(-)in the spreadsheet
represents relational info	(-)represents contextual info
"knowledge Tool"	**"data Tool"**

Here, the black entries are more salient than the gray ones (cited for clarification).

Except for the implicit pole "outside of spreadsheet" all others clearly refer to the information sources to give access to KNOWLEDGE. In contrast, the implicit poles indicate that the elements turn given DATA into INFORMATION. The outlier construct "outside/inside spreadsheet" can be explained by the high concentration of `SX` objects placed near this pole and that `SX` was considered an add-on. The content of information was categorized. The elements were rated as tools to provide a specific kind of information. Therefore, we tag the PC_1 dimension ranging from **"data Tool"** to "**knowledge Tool**".

3.0.2 Principal Component PC_2

The second component (vertical in Fig. 1) can be described best by the following constructs:

Implicit Pole	**Emergent Pole**
visual information	(-)cognitive information
project-specific meaning	(-)globally defined meaning
(-)super category	more specific category
(-)implicit meaning	overt meaning
concrete relation to spsht	(-)location-indep. info
(-)pure info	functional info
"Represented data"	**"Implicit knowledge"**

The distribution of implicit and emergent poles with respect to PC_2 wasn't uniform. The poles of the two most salient constructs agreed, so we called the resp. PC_2 construct poles accordingly. Note that "meaning" in these pole names does not refer to importance but rather to denotation. Also, the construct "super versus more specific category" is ambiguous: the subject wanted to distinguish between a higher-level, hidden and a lower-level, but explicit structural aspect of the information conveyed by the elements.

All constructs are more concerned with rating the communicated information itself, thus we can see a differentiation along Probst' information traits from DATA to KNOWLEDGE. Moreover, the interviewees distinguished the degree of direct recognition of structure ranging from explicit representation to implicit context of information. Therefore, we label the PC_2 dimension with **"Represented data**" and "**Implicit knowledge**".

3.0.3 Principal Component PC_3

In analogy, we determine the third principal component description by analyzing the constructs closest to PC_3 in the biplot of PC_3 against PC_1 resp. PC_2. The same constructs were most salient at a suppression of 0.8% in both biplots:

Implicit Pole	**Emergent Pole**
info for the author	(-)info for the reader
presentation	(-)computation
help creating spreadsheets	(-)help understanding
concrete info	(-)abstract info
generating data	(-)exploring data
"Author"	**"Reader"**

PC_3 positively distinguishes the elements according to their purpose when used by spreadsheet authors versus readers.

3.1 Discussion

Let us now combine the findings about the PC constructs and element clusters. Fig. 4 summarizes the findings, which we will discuss in the subsequent paragraphs. Note that the term "versus" in the subtitles does not signify opposition, but is supposed to enhance readers' distinct context experiences implied by our subjects' construct elicitations.

3.1.1 Information Perception versus Interface Perception

Fig. 4 visualizes the element distribution according to the Principal Component Analysis as in Fig. 1. The only difference is that we enhanced the distance between the element clusters to allow for a horizontal grid to depict the distinctions discussed in the following.

Figure 4: Interpretation of Fig. 1

The first principal component dimension ranges from "DATA Tool" to "KNOWLEDGE Tool", hence we use the knowledge management model components GLYPH, DATA, INFORMATION, and KNOWLEDGE (Fig. 3) as scale. The exact location of these on the x-axis of Fig. 4 is determined by observing the specific transformation function of the spreadsheet player's information objects in terms of the model. A result of our investigation is the recognition that spreadsheet readers rate interface information objects according to their respective cognitive quality: Does an interface object carrying information support DATA, INFORMATION, or KNOWLEDGE gain? An evaluation scheme for information objects may thus be based on the question what information trait they offer for the reader.

In contrast, the second most relevant aspect under which information objects are perceived is given by the second principal component construct "Represented DATA— Implicit KNOWLEDGE". Here, they are rated for the information quality itself. Is, for instance, the information communicated explicitly as DATA, e.g. by using explicit second notation, or is it given as KNOWLEDGE, e.g. by showing dependencies based on background knowledge assumptions?

Thus, spreadsheet readers distinguish the cognitive information conditioning from the concrete information offering. For example, the elements `Tables` and `sx:Localized Info` are rated similarly with respect to their framing as tools that enable a reader to get INFORMATION. But they were evaluated very differently regarding their entropy, the former is considered to represent INFORMATION as DATA, whereas the latter contains implicit KNOWLEDGE.

3.1.2 A data-to-information versus information-to-knowledge Environment

If we look in Fig. 1 at the element space with the coordinate system slightly shifted, then *all* common spreadsheet information objects are on the left side and *all* `SX` ones are on the right side (except for the hybrid `(sx:)Color Coding` which is located very close to the separating axis). In particular, the PC_2 dimension "KNOWLEDGE Tool – DATA Tool" separates the one set of elements from the other. The information services offered by the `SX` extension and common spreadsheet applications as perceived by readers do not overlap. If we agree with the interpretation in Fig. 4 we can even acknowledge a progression between these element sets, where the common set serves as a DATA-to-INFORMATION interface, whereas the `SX` set provides a INFORMATION-to-KNOWLEDGE interface.

Poles	Elements
"Author"	`Formulae`, `Values`, `Tables`, `sx:Functional Block`
"Reader"	`sx:Dependency Graph`, `sx:Relational Arrows`, `Legends`, `sx:Localized Info`, `sx:Concept Nodes`, `Headers`, `Title`, `(sx:)Color Coding`

3.1.3 Spreadsheet Authors' versus Readers' Information Sources

Information is important for spreadsheet authors during the creation process and-for readers in the interpretation process. The analysis of the third principal component PC_3 disclosed that readers differentiate information objects in spreadsheets according to their use by authors or readers. But which elements are for whom?

According to the PC_1-PC_3- as well as the PC_2-PC_3 biplot the elements are consistently in the resp. halfspace as to the distinction between author and reader use:
The order of elements is determined by their decreasing mapping relating to the respective pole. Interestingly, this indicates that readers do consider only formulae (calculation), values (database data), tables(database views) and block arrangements (structural design) as creative choice options for the author. All the other information is intrinsically determined by the other elements. This suggests automatisation tasks for spreadsheet applications of the future.

3.1.4 Inside Objects versus Outside Objects in Spreadsheets

In each element cluster per definitionem elements were rated similarly by our interviewees. Nevertheless, it is striking that the clusters are located very distinctively with not only no intersection but also with a significant margin inbetween them. Basically each cluster inhabits a quadrant of Fig. 1 by itself. Looking at the elements closely, we note that cluster 2 except for `Title` contains all elements which refer to the information for one cell. They not only take this micro-perspective, their position is also close to the resp. cell. Cluster 1, in contrast, comprises all information objects that are concerned with a connected range of cells, we can say a meso-perspective. Finally, cluster 3 members show domain-oriented information, a macro-perspective so to speak. Additionally, these information objects are located next to the spreadsheet application, particularly not within the application. This correlation of places and information framings is not random as interviewees used the constructs consistently together. It should be researched whether the position on screen provides a usability issue for the `SX` extension.

3.1.5 The Desktop versus a Communication Metaphor

In office suites the desktop metaphor prevails: Documents (like text files) are managed and can be accessed via players (like text processors). A **player** "plays" data, whereas a **document** "documents" data. The distinction between information

objects with respect to being an interface tool (PC_1) versus information representation (PC_2) can be interpreted as referring to player-dependent and document-dependent properties.

For spreadsheets this distinction is rather surprising, as neither spreadsheet *programs*(!) are typical documents nor spreadsheet *applications*(!) typical players. Programs transform input data into output data, application software is "computer software designed to help the user to perform specific tasks". We noted before that the spreadsheet author view is predominant when looking at spreadsheets in research. Our terminology turns out to be yet another proof. Therefore, it is important to note that readers seem to distinguish between the document and player line. Clarification discussions with some of the interviewees revealed that this distinction was not an explicit one. Prompted to differentiate between the two, the subjects were surprised and not able to distinguish the concepts continuously. They experience a spreadsheet program together with the application as an entity. This might be due to the inconsistencies of the document/player metaphor for spreadsheets: Does the computational execution of a formula belong to the document or the application?

The integration of spreadsheets into office suites and the force of the underlying metaphor carries the consequence that only the individual (authored) data are saved into files, that in turn are exchanged like documents. But spreadsheet content, that is, basically numbers, is much more context-dependent than e.g. natural language content. Therefore, we suggest to switch metaphors: For readers spreadsheets serve essentially as a communication-of-information tool. If we took up the communication metaphor, then we need to think about context-integration on all levels: Which context must be distributed along with the file data of today? Which context presentation schemes can be employed in the application? For example, it is well-known that mathematicians have improved mathematical formats over hundred of years to obtain a visual language for symbols that conveys more context to a reader than common text. Spreadsheets ignore this specific interpretation help for now.

4 Conclusion

Interpretation and comprehension of spreadsheets constitute a rather neglected usability issue in research concerned with spreadsheets. In this paper, we presented a repertory grid study and subsequent General Procrustes Analysis that explore qualitative properties of information objects in spreadsheets from the point of view of spreadsheet readers. We discussed five framings of information sources in spreadsheets that readers perceived:

Perception Dimensions Our interviewees perceived information objects differently when considered as tools or with respect to their information content. As tools, e.g., two sources can offer the same kind of information (`Tables`

and `sx:Localized Info` provide INFORMATION), but with respect to their content they can be rated quite differently (`Tables` represents information as DATA whereas `sx:Localized Info` provides it as KNOWLEDGE).

Role-Specificity Our investigation showed certain information objects strongly associated with authors vs. readers.

Information Environment The set of common spreadsheet information sources (as for example offered by `MS Excel`) create an environment that enables turning DATA into INFORMATION. In contrast, the exemplaric set of `SX` information sources were identified as an environment that enables turning INFORMATION into KNOWLEDGE.

Neighborhood of Information The position of information sources inside or outside the frame of the application was observed by our interviewees. Elements were even clustered according to the position of their respective point of reference.

Metaphoric Boundaries Readers' distinction between document-dependent and player-dependent information is traced to the underlying desktop metaphor. But for spreadsheets this metaphor is rather limiting since the context-dependancy of numbers is neglected. Moreover, for readers the document/player metaphor is also restrictive as from their perspective the main purpose of spreadsheets consists in their communication, not in their documentation functionality.

These framings of information from the readers' point of view represent relations between the set of information objects, hence we have a first readers' information model of spreadsheets. As our study was only an exploratory one, we cannot conclude this information model to be complete, nor can we evaluate the ranking of different framings in general. But it has become clear by this study that readers have their own interesting perspective on information offered in spreadsheets. In the near future we want to design another repertory grid study which will use not only a fixed set of elements but also a fixed set of representative constructs (based on the constructs elicited here). Such an RGI can be automated and for example exposed to a crowdsourcing Internet marketplace as the Amazon Mechanical Turk to get significant statistical data for general results.

Finally, based on the found framings we like to suggest new interactions to increase the usability of spreadsheets especially for readers:

- The *perception of distinct dimensions of information objects* points to a frequently neglected media-theoretic topic that also concerns spreadsheets: Information objects are media and as such they do not only contain a message, they also are the message [19]. When using e.g. the information object `Tables`, then input data are perceived as DATA by readers. As DATA they need a context to become meaningful, but at the same time `Tables` as a

structured, formal notation carries a 'truth' statement. Therefore, readers trust the information they get, even though the information object itself delivers no context to turn the DATA into INFORMATION. As a consequence authors should be compelled to create context, e.g. respective `Headers` or `Legends` if the spreadsheet is meant to be distributed, and readers should be required to understand the context before interpreting the DATA. The former is realized in many spreadsheet extensions/applications already, but the latter is not.

- The perceived differentiation of spreadsheet users into *authors and readers* allows a much better fine-tuning of services. Even though the existence of both groups has been recognized, the interface design for players has not yet seriously taken this distinction into account. We can think of more role-specific information services for readers. If readers, for instance, want to understand specific parts of a spreadsheet, these parts could be rendered separately on the fly as a spreadsheet view. This can reduce the cognitive overload when interpreting numbers in a big spreadsheet, particularly if information is scattered over multiple worksheets. Another reader specific service consists of a better navigation within spreadsheets, e.g. a semantically driven navigation as already prototypically presented with CogMap [11] or with `SX`' semantic navigation [16].

- Our interviewees distinguished between *information environments* that turn DATA into INFORMATION and ones that turn INFORMATION into KNOWLEDGE. In other words, they considered typical spreadsheet applications like `OpenOffice` or `MS Excel` as data interfaces, whereas the extension `SX` was considered a knowledge interface.[4] This induces the question how we can further enhance a data interface with "meta level" information objects. For instance, we could provide a reader access to the provenance of data or we could help the reader to assess information.

- Following a *communication metaphor* for a reader, a communication mode of spreadsheets can be enabled, that provides on the one hand access to *document-specific* experts, background ontologies, or fora and on the other hand access to *topic-driven* discussions, domain knowledge e.g. in standard financial text books, help fora, or other domain services.

- If we set the *document/player metaphor aside* then spreadsheets can also make use of already developed, open-standard, but non-spreadsheet-specific format guidelines: mathematical formatting of formulae.[5] For this, we can

[4] We consider the discriminatory power of this distinction a consequence of not having included common analysis support tools of spreadsheets like diagrams for our element set.

[5] We are fully aware that this might not be the best for spreadsheet author, see e.g. [20], even though we suggest to think about it for complex formulae as well. For such, in our opinion, the typical spreadsheet formula language is not visual enough.

imagine a math editor and viewer, which takes input e.g. in LaTeX form – commonly used by mathematicians (which are typically non-programmers) for writing complex formulae –, converts it into MathML and renders it for reading in a browser window in standard mathematical notation. As our study indicated a neighborhood-of-information framing, we envision the window to be close to the cell for which such a formula is created.

Note that many of the envisioned interactions may be generalized to other office suite members to improve readability and ,thus, usability.

All in all, we believe that the framings of information in spreadsheets by readers presented in this paper are the entry door for a better, more complete understanding of human-spreadsheet interaction and a new source for according design inspirations.

References

[1] Kenneth R. Baker et al. "Comparison of Characteristics and Practices amongst Spreadsheet Users with Different Levels of Experience". In: *CoRR* abs/0803.0168 (2008).

[2] Jonathan P. Caulkins, Erica Layne Morrison, and Timothy Weidemann. "Spreadsheet Errors and Decision Making: Evidence from Field Interviews". In: *JOEUC* 19.3 (2007), pp. 1–23.

[3] Chris Chambers and Chris Scaffidi. "Struggling to Excel: A Field Study of Challenges Faced by Spreadsheet Users". In: *Proceedings of the 2010 IEEE Symposium on Visual Languages and Human-Centric Computing*. VLHCC '10. Washington, DC, USA: IEEE Computer Society, 2010, pp. 187–194. ISBN: 978-0-7695-4206-5.

[4] Sarah E. Crudge and Frances C. Johnson. "Using the information seeker to elicit construct models for search engine evaluation". In: *JASIST* 55.9 (2004), pp. 794–806.

[5] EUSPRIG. *European Spreadsheet Risks Interest Group*. home page at `http://www.eusprig.org`. 2010. URL: `http://www.eusprig.org`.

[6] J. Gower. "Generalized procrustes analysis". In: *Psychometrika* 40 (1 1975), pp. 33–51. ISSN: 0033-3123.

[7] T. R. G. Green and M. Petre. "Usability Analysis of Visual Programming Environments: a 'cognitive dimensions' framework". In: *JOURNAL OF VISUAL LANGUAGES AND COMPUTING* 7 (1996), pp. 131–174.

[8] James W. Grice. *Generalized Procrustes Analysis Example with Annotation*. `http://www.idiogrid.com/GPA_Idiogrid_Example.pdf`. 2007.

[9] James W. Grice. “Idiogrid: Software for the management and analysis of repertory grids”. In: *Behavior Research Methods, Instruments, & Computers* 34 (2002), pp. 338 –341.

[10] James W. Grice and Kimberley K. Assad. “General Procrustes Analysis: A Tool for Exploring Aggregates and Persons”. In: *Applied Multivariate Research* 13.1 (2009), pp. 93–112.

[11] David G. Hendry and Thomas R. G. Green. “CogMap: a Visual Description Language for Spreadsheets”. In: *J. Vis. Lang. Comput.* 4.1 (1993), pp. 35–54.

[12] David G. Hendry and Thomas R. G. Green. “Creating, comprehending and explaining spreadsheets: a cognitive interpretation of what discretionary users think of the spreadsheet model”. In: *Int. J. Hum.-Comput. Stud.* 40.6 (1994), pp. 1033–1065.

[13] Devi Jankowicz. *The Easy Guide to Repertory Grids*. Wiley, 2003. ISBN: 0470854049.

[14] George Kelly. “International Handbook of Personal Construct Technology”. In: John Wiley & Sons, 2003. Chap. A Brief Introduction to Personal Construct Theory, pp. 3–20.

[15] Andrew J. Ko et al. “The state of the art in end-user software engineering”. In: *ACM Comput. Surv.* 43.3 (Apr. 2011), 21:1–21:44. ISSN: 0360-0300.

[16] Andrea Kohlhase. “Towards User Assistance for Documents via Interactional Semantic Technology”. In: *KI 2010: Advances in Artificial Intelligence*. Ed. by Rüdiger Dillmann et al. LNAI 6359. Karlsruhe, Germany, 2010, pp. 107–115.

[17] Clayton Lewis and Gary Olson. “Can principles of cognition lower the barriers to programming?” In: *Empirical studies of programmers: Second workshop*. Ed. by Gary M. Olson, Sylvia Sheppard, and Elliot Soloway. Empirical studies of programmers. Norwood, NJ, USA: Ablex Publishing Corp., 1987, pp. 248–263.

[18] Cliff McKnight. “The personal construction of information space”. In: *Journal of the American Society for Information Science* 51.8 (2000), pp. 730–733. ISSN: 1097-4571.

[19] Marshall McLuhan. *Understanding media: The extensions of man*. New York: McGraw-Hill, 1964.

[20] Bonnie A. Nardi. *A Small Matter of Programming: Perspectives on End User Computing*. 1st. Cambridge, MA, USA: MIT Press, 1993. ISBN: 0262140535.

[21] Bonnie A. Nardi and James R. Miller. “An Ethnographic Study of Distributed Problem Solving in Spreadsheet Development”. In: ACM Press, 1990, pp. 197–208.

[22] Bonnie A. Nardi and James R. Miller. "The spreadsheet interface: A basis for end user programming". In: *Proceedings of the IFIP TC13 Third Interational Conference on Human-Computer Interaction*. INTERACT '90. Amsterdam, The Netherlands, The Netherlands: North-Holland Publishing Co., 1990, pp. 977–983. ISBN: 0-444-88817-9.

[23] Gregory B. Newby. "Cognitive space and information space". In: *JASIST* 52.12 (2001), pp. 1026–1048.

[24] Raymond R. Panko. "Spreadsheet Errors: What We Know. What We Think We Can Do." In: *Symp. of the European Spreadsheet Risks Interest Group (EuSpRIG 2000)*. 2000.

[25] Raymond R. Panko. "What we know about spreadsheet errors". In: *Journal of End User Computing* 10 (1998), pp. 15–21.

[26] Stephen G. Powell, Kenneth R. Baker, and Barry Lawson. "A critical review of the literature on spreadsheet errors". In: *Decision Support Systems* 46.1 (2008), pp. 128–138.

[27] Stephen G. Powell, Barry Lawson, and Kenneth R. Baker. "Impact of Errors in Operational Spreadsheets". In: *CoRR* abs/0801.0715 (2008).

[28] G. Probst, St. Raub, and Kai Romhardt. *Wissen managen*. 4 (2003). Gabler Verlag, 1997.

[29] Christopher Scaffidi, Mary Shaw, and Brad A. Myers. "Estimating the Numbers of End Users and End User Programmers". In: *VL/HCC*. 2005, pp. 207–214.

[30] Felix B. Tan and M. Gordon Hunter. "The Repertory Grid Technique: A Method for the Study of Cognition in Information Systems". English. In: *MIS Quarterly* 26.1 (2002), pp. 39–57. ISSN: 02767783.

[31] K. Wolstencroft et al. "RightField: Embedding ontology annotation in spreadsheets". In: *Bioinformatics* 24.14 (2011), pp. 2021–2022.

Andrea Kohlhase has a graduate degree in Mathematics and a PhD in Computer Science. She is interested in the intersection of Semantic Technologies and Interaction Design: What new interactions can evolve with Semantic Technologies? At the moment she works as a researcher at Jacobs University Bremen and as a lecturer of Human Factors at the University of Applied Sciences in Bremen.