Human-Spreadsheet Interaction

Andrea Kohlhase

Jacobs University Bremen, School of Engineering and Sciences, Germany
a.kohlhase@jacobs-university.de

Abstract. Spreadsheets have become very popular tools for analyzing and visualizing data from business and science. To better understand human-spreadsheet interaction, we explore readers' information models, but in contrast to most studies we focus on spreadsheet readers rather than spreadsheet authors. We conducted a study using the repertory grid technique and analyzed the result with the help of a Generalized Procrustes Analysis yielding a deeper understanding of human's information model of spreadsheets. Based on this we envision new human-spreadsheet interactions to increase the readibility and thus, usability of spreadsheets.

Keywords: Spreadsheets; repertory grid; information model of spreadsheets; human-spreadsheet interaction; information objects

1 Introduction

The intuitive, flexible, and direct approach to computation in spreadsheets has led to widespread use and reuse. In particular, spreadsheets 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 [16]. This intensity yields not only more and more shared, complex spreadsheet programs, but also high-impact errors on the data level (up to 90% are affected according to [16]) and on the comprehension level (see [17]). The losses caused by formula errors and misinterpretation have even led to an international task force to battle them [4].

Previous Approaches The standard research addressing such usability problems is based on Panko's influential report on error states and types in [16], that mainly considers the data level, e.g., computational errors based on faulty formulae.

Lewis and Olson gave a critical cognitive psychology account for the success of spreadsheets [12]. 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. Consequently, an abundance of research was directed towards end-user programming (EUP), i.e., "programming to achieve the result of a program primarily for personal […] use" [10, p. 4]. Unfortunately, from the point of view of

EUP, a spreadsheet user is reduced to an end-user, that is simply any computer user who creates a spreadsheet program.

Spreadsheet Readers as Spreadsheet Users But Nardi and Miller noticed another feature of spreadsheets in [15]: they are not "single-user applications". In particular, they are used in the work environment on the one hand as collaboration tool on the other as means of communication to exchange and combine domain knowledge and programming expertise. In [7] Hendry and Green highlight the fact that spreadsheet use is also a matter of understanding. They report their informants' missing comprehension (even of their own spreadsheet) and trace it to lacking comprehensibility support by spreadsheets, i.e., a specific usability issue for people reading spreadsheets. Such "readers" are, for instance, people that make use of existing templates by simply putting in new data, review data developments on different abstraction levels (like supervisors), assess data to base further decisions upon, want to understand their own spreadsheet program after a while, or look for reusable parts of a spreadsheet program, therefore browsing available ones. This means that an essential part of the spreadsheet user base is not yet addressed by current research. Many reports about bad decisions caused by misinterpretation and difficulties of spreadsheet comprehension confirm this observation (e.g. [2, 1, 19]).

Our Approach Sometimes it is said that spreadsheets can be considered *data interfaces* that display and allow to play with data. Note though that data by itself is not interesting. In particular, Probst et al. suggested in [18] a knowledge management model. It differentiates 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 question, in how far readers really consider spreadsheets merely as DATA interfaces, is the starting point for this paper. Moreover, we were interested in the evaluation criteria of readers with respect to a spreadsheet's information value. Based on such an analysis we are able to understand more deeply the information model of spreadsheets and can suggest innovative human-spreadsheet interactions.

Methodology: Repertory Grids and General Procrustes Analysis Spreadsheet readers distinguish interface objects that carry information ("**information objects**"). 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** [9, 8]. RGI explores personal constructs, i.e., how persons perceive and understand the world around them. McKnight was the first to suggest RGI as a semi-empirical method for exploration of an information space [13]. By now RGI is a well-established method to explore users' personal constructs when interacting with software artifacts (see [21] for a list of examples). A crucial advantage over other methods is that a small sample size can be used.

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. 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.

We analyzed the repertory grid data obtained in the main study with "Idiogrid" and followed the analysis as described by Grice in [5]. In particular, we performed Gower's "**Generalized Procrustes Analysis (GPA)**".

2 The RGI Study[1]

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 (see Table 1). Note that "diagrams" are missing, which may be due to the fact, that they were not part of our standard spreadsheet example. To broaden this set of elements, we added information objects which are not traditionally used in spreadsheets (see Table 2). 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 semantic spreadsheet extension "**SACHS**" [11]. The union of both sets of information objects were used as the given, fixed set of elements, for which in this RGI study constructs were to be elicited by the interviewees.

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:)ColorCoding`	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

Interview Procedure In the study we presented each subject a simple, but complexly structured spreadsheet. Each element was explained by the interviewer and SACHS was introduced where necessary. Following traditional RGI, the interviewee was then

[1] A full description can be found at http://jpubs.jacobs-university.de/handle/579/2453.

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?

Table 2. Extra Information Objects of Spreadsheets

`sx:LocalizedInfo`	A local look-up (data and text) of relevant information for cells on a by-cell-click basis
`sx:FunctionalBlock`	A local border indicating all cells functionally associated to the currently selected cell
`sx:DependencyGraph`	An overview graph (in a different window) of concepts showing on which the corresponding (selected) cell is ontologically dependent
`sx:RelationalArrows`	An arrow indicating a dependency relation between concepts in `sx:DependencyGraph`
`sx:ConceptNodes`	A node in `sx:DependencyGraph` representing a dependent subconcept, that additionally serves as a link to corresponding spreadsheet cells

Interview Summary For our investigation we interviewed 14 people, of which 10 were male and 4 female. The age distribution was the following: 5 persons were under the age of 20, 6 between 20 and 30, 2 between 30 and 40, and 1 between 40 and 50. 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 `SACHS` before the interview. The rating scale was binary. 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.

The first component of our GPA returned a rather high similarity (.68%). It was tested for statistical significance with the help of a randomization test based on 500 trials, which yielded an observed proportion $p \leq 0.00$ - verifying statistical significance. A subsequent standard **Principal Components Analysis (PCA)** produced $\{PC_{1\ldots11}\}$. The first component explains ca 33.7%, the second 22.2% and the third 14.4% of the variance in the data. To approximate the meaning of the PCs, we looked at elicited similar – i.e., more salient (84%) – constructs and determined dimensions that can serve as common denominator constructs. As a result the PC_1 dimension is interpreted to range from “Data Tool” to “Knowledge Tool” (interface perception), PC_2 aligns to “Represented Data— Implicit Knowledge” (info perception), and PC_3 differentiates between spreadsheet use by authors or readers. The reliability of this qualitative analysis was ensured by following the procedure given in [8, 155ff.].

2.1 Towards an Information Model of Spreadsheets

Fig. 2 visualizes the element distribution according to the PCA, which we will discuss in the following. The x-axis corresponds to the interpretation of PC_1 as readers' perception of the interface, the y-axis to PC_2 as their perception of information.

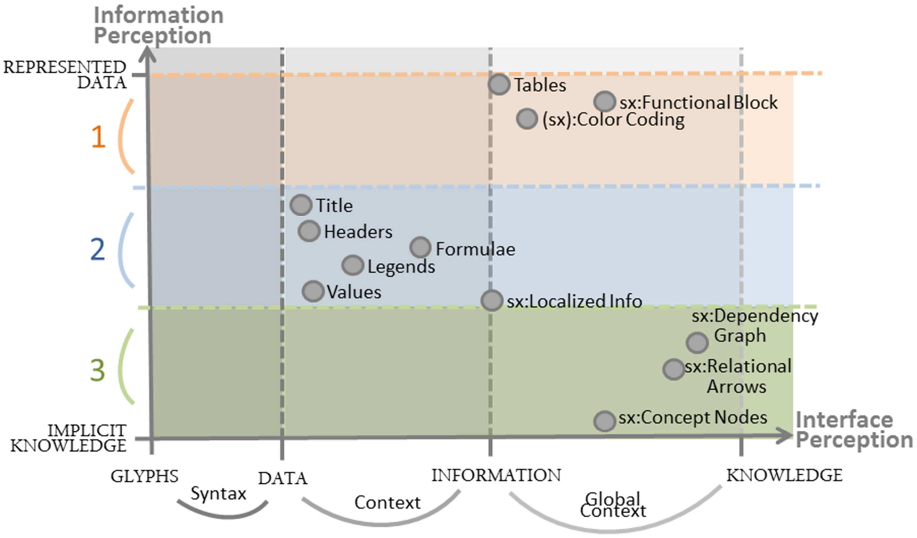

Fig. 1. Interpretation of Results

Perception Dimensions Our interviewees perceived information objects differently when considered as tools (interface perception) or with respect to their information content (information perception). 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. Interestingly, this indicates that readers only consider formulae (calculation), values (database data), tables (database views) and block arrangements (structural design) as creative options for the author. All the other information is intrinsically determined by the other elements. This suggests automation opportunities for spreadsheet applications of the future.

Information Environment The information services offered by the `SACHS` extension and common spreadsheet applications do not overlap in the perception of readers. We can even acknowledge a progression between these element sets, where the common element set from Table 1 serves as a DATA-to-INFORMATION interface, whereas the `SACHS` set from Table 2 provides an INFORMATION-to-KNOWLEDGE interface (see Fig. 2).

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

Metaphoric Boundaries 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. For spreadsheets this very distinction by readers is rather surprising, as neither spreadsheet *programs*(!) are typical documents nor spreadsheet *applications*(!) typical players. Moreover, the desktop metaphor is rather limiting For spreadsheets, since the context-dependency of numbers is neglected. Finally, 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.

3 Innovative Human-Spreadsheet Interactions

Based on the found dimensions we like to suggest new interactions to increase the usability of spreadsheets especially for readers:

1. 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 [14]. 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.
2. 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 will need 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 [6] or with `SACHS`' semantic navigation [11].
3. Our interviewees distinguished between *information environments* that turn DATA into INFORMATION and ones that turn INFORMATION into KNOWLEDGE. 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.
4. 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.

5. 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.[2] For this, we can 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, that is, the communication aspect.

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: Interface vs. information perception,

Role-Specificity: Information objects for authors vs. readers,

Information Environment: From DATA to INFORMATION with e.g. `MS Excel`, but from INFORMATION to KNOWLEDGE with e.g. `SACHS`,

Neighborhood of Information: Positioning of information sources inside or outside the frame of the application, and

Metaphoric Boundaries: A desktop vs. communication metaphor.

These dimensions of information from the readers' point of view represent relations between the set of information objects. Thus, they can serve as a first information model of spreadsheets and based on this we envisioned some innovative forms of human-spreadsheet interaction in the last section. As our study was only exploratory, we cannot conclude this information model to be complete. Nevertheless, it has become clear by this study that readers have their own interesting perspective on information offered in spreadsheets.

All in all, the information model of spreadsheets (by readers) presented in this paper is the entry door for a better, more complete understanding of human-spreadsheet interaction and a new source for according design inspirations.

[2] We are fully aware that this might not be the best for a spreadsheet author, even though for complex formulae the typical spreadsheet formula language is not visual enough.

Acknowledgements This work has been funded by the German Research Council under grant KO-2484-10-1.

5 References

1. J. P. Caulkins, E. L. Morrison and T. Weidemann (2007) Spreadsheet errors and decision making: evidence from field interviews, *JOEUC* 19 (3), pp. 1–23.
2. C. Chambers and C. Scaffidi (2010) Struggling to excel: a field study of challenges faced by spreadsheet users, VLHCC ’10, Washington, DC, USA, pp. 187–194. ISBN 978-0-7695-4206-5
3. R. Dillmann, J. Beyerer, U. D. Hanebeck and T. Schultz (Eds.) (2010) Proceedings of the 33.rd annual german conference on artificial intelligence ki’10, LNAI, Karlsruhe, Germany.
4. EUSPRIG (2010) European spreadsheet risks interest group, http://www.eusprig.org
5. J. W. Grice (2007) Generalized procrustes analysis example with annotation,
6. D. G. Hendry and T. R. G. Green (1993) CogMap: a visual description language for spreadsheets, *J. Vis. Lang. Comput.* 4 (1), pp. 35–54.
7. D. G. Hendry and T. R. G. Green (1994) Creating, comprehending and explaining spreadsheets: a cognitive interpretation of what discretionary users think of the spreadsheet model, *Int. J. Hum.-Comput. Stud.* 40 (6), pp. 1033–1065.
8. D. Jankowicz (2003) The easy guide to repertory grids, Wiley. ISBN 0470854049
9. G. Kelly (2003) International handbook of personal construct technology, pp. 3–20.
10. A. J. Ko, R. Abraham, L. Beckwith, A. Blackwell, M. Burnett, M. Erwig, C. Scaffidi, J. Lawrance, H. Lieberman, B. Myers, M. B. Rosson, G. Rothermel, M. Shaw and S. Wiedenbeck (2011-04) The state of the art in end-user software engineering, *ACM Comput. Surv.* 43 (3), pp. 21:1–21:44. ISSN 0360-0300
11. A. Kohlhase (2010) Towards user assistance for documents via interactional semantic technology, in [3], pp. 107–115.
12. C. Lewis and G. Olson (1987) Can principles of cognition lower the barriers to programming?, (S. Sheppard and E. Soloway Eds.), Empirical studies of programmers, Norwood, NJ, USA, pp. 248–263.
13. C. McKnight (2000) The personal construction of information space, *Journal of the American Society for Information Science* 51 (8), pp. 730–733. ISSN 1097-4571
14. M. McLuhan (1964) Understanding media: the extensions of man, McGraw-Hill, New York.
15. B. A. Nardi and J. R. Miller (1990) An ethnographic study of distributed problem solving in spreadsheet development, pp. 197–208.
16. R. R. Panko (2000) Spreadsheet errors: what we know. what we think we can do., in [20],
17. S. G. Powell, K. R. Baker and B. Lawson (2008) A critical review of the literature on spreadsheet errors, *Decision Support Systems* 46 (1), pp. 128–138.
18. G. Probst, St. Raub and K. Romhardt (1997) Wissen managen, 4 (2003) edition, Gabler Verlag.
19. C. Scaffidi, M. Shaw and B. A. Myers (2005) Estimating the numbers of end users and end user programmers, pp. 207–214.
20. (2000) Symp. of the european spreadsheet risks interest group (eusprig 2000),
21. F. B. Tan and M. G. Hunter (2002) The repertory grid technique: a method for the study of cognition in information systems, *MIS Quarterly* 26 (1), pp. pp. 39–57.