A FRAMEWORK FOR THE IDENTIFICATION OF SPREADSHEET USAGE PATTERNS

Complete Research

Reschenhofer, Thomas, Technical University of Munich, Munich, Germany, reschenh@in.tum.de
Matthes, Florian, Technical University of Munich, Munich, Germany, matthes@in.tum.de

Abstract

Spreadsheets are omnipresent in today's enterprises, in particular for supporting business users in their decision-making. Thereby, they are used in various ways to support business processes, i.a., in the areas of financial reporting, resource management, and project management. However, shortcomings of prevalent spreadsheet software (e.g., limited collaboration support) have negative impacts onto the risk and compliance efforts which are dictated by both the company's goal of reducing the risk of financial losses as well as by legal regulations. Due to the variety of spreadsheet usages, there cannot be single solution addressing all those shortcomings. It is necessary to define classes of spreadsheet usages in order to be able to develop tailored solutions for each of those classes.
Therefore, in this paper we show the results of an empirical study on spreadsheet usages, which we conducted in two companies. Based on those results we propose a morphological box as a classification framework for determining patterns of spreadsheet usages. Those patterns can serve as a foundation for future research focusing on specific spreadsheet usage patterns, e.g., the design of tailored solution approaches for enhancing the support of business processes.

Keywords: Spreadsheet, Usage Pattern, Empirical Study, Morphological Box.

1 Introduction

In today's enterprises, spreadsheets as decision-support tools are indispensable (Panko, 2008). They are used for various purposes, e.g., financial reporting (Panko, 2006), business analysis (Winston, 2001), or workload planning (Pemberton and Robson, 2000). Bradley and McDaid (2009) state that at least 90% of all desktops have spreadsheets installed. Because of the overwhelming quantity and importance of spreadsheets, it is no surprise that Panko and Port (2012) are calling spreadsheets—and end user computing (Nardi, 1993) in general—the "dark matter" of corporate information technology. As the dissemination of spreadsheets increased over the last decades, numerous studies have shown that most of them contain errors (Panko, 2006), whereas not few of those errors led to significant financial losses (Caulkins et al., 2007; Powell et al., 2009). Hence, lots of research was done to study and classify spreadsheet errors (Powell et al., 2008) and to develop methodologies and approaches to reduce the risk and impact of errors in spreadsheets (Cunha et al., 2012; Engels and Erwig, 2005; Hermans, 2012; Janvrin and Morrison, 2000). While errors of spreadsheets were already observed extensively (Powell et al., 2008), there is still very little research about spreadsheet shortcomings from an Information System (IS) research perspective (Braa and Vidgen, 1999; Davis and Olson, 1985) taking into account all three of people, processes, and technology. Therefore it is still very unclear, how good spreadsheets usually support respective business processes (e.g., financial reporting). By looking at spreadsheets from an IS research perspective, they still might suffer from shortcomings with respect to their usage as part of a specific business process, even if the spreadsheets as such are completely free of errors. For example, shortcomings like the restriction to manual

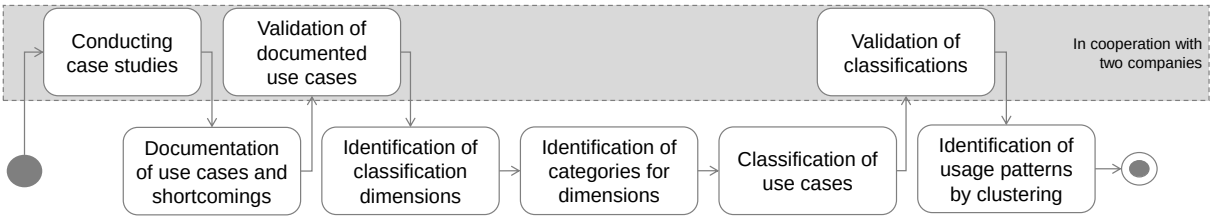

Figure 1: An illustration of this work's research approach.

processes, limited collaboration support, and lacking scalability significantly impede an enterprise's risk and compliance efforts (MetricStream GRC Blog, 2014). Moreover, not only the endeavor of reducing the risks induced by the application of prevalent spreadsheets, but also legal regulations like the *Sarbanes-Oxley Act* (2002) in the US or *Basel II* (2004) in the EU are forcing enterprises to reevaluate and enhance the application of spreadsheets for the support of certain business processes.

Since spreadsheets are applied for a plethora of purposes (Ronen et al., 1989), there will be no general solution enhancing the support of all diverse business processes. Hence there have to be specific solutions for certain spreadsheet usage patterns with respect to the respective business process support, since many shortcomings are related to the actual type of usage of the spreadsheet (Grossman et al., 2005). Therefore a tailored solution can address the specific shortcomings of a spreadsheet usage pattern while neglecting the specific issues of others. However, so far there is little research about usage patterns of spreadsheets. The present work presents an approach for the determination of spreadsheet usage patterns by proposing a classification framework for spreadsheets. Based on this, patterns can be identified by clustering spreadsheets with similar usage characteristics according to the proposed classification scheme. Therefore, the answers to the following research questions constitute this work's contribution:

1. What is a suitable framework for the determination of spreadsheet usage patterns?
2. What are typical spreadsheet usage patterns in industry?

Based on the proposed classification framework as well as the identified usage patterns, future research activities can focus on enhancements of spreadsheet solutions for specific spreadsheet patterns and address specific shortcomings of spreadsheets within this patterns.

The remainder of this paper is organized as follows: Section 2 summarizes related work, in particular about spreadsheet errors, user studies about problems of spreadsheets, and spreadsheet applications in business. The first phase of the present works research approach was the identification and observation of spreadsheets in practice. By applying explorative case research as defined by Benbasat et al. (1987) and Yin (2014) in two companies, we studied the design, scope, usage, and context of nine spreadsheet applications. The setting and the outcomes of these studies are described in Section 3. Based on those use cases and on related literature, Section 4 then describes the derivation of a classification framework consisting of seven dimensions with respective characteristics in order to be able to identify cross-case patterns (Eisenhardt, 1989). Thereafter, Section 5 describes the identification of spreadsheet patterns based on the observed use cases and the developed classification framework. Figure 1 illustrates the research approach, whereas the approach and the results of this work are critically reflected in Section 6. Drawing on the results presented in this paper, Section 7 addresses the research questions presented above and discusses implication for future research.

2 Related Work

Spreadsheets have been subject of research for more than 25 years (Panko, 2006). In the following section, we summarize related activities in the discipline of spreadsheet research. By orienting on the IS research perspective (people, process, technology), we focus on the research of spreadsheet errors in general (technology), the role of the user (people), and the application of spreadsheets in business (process).

Research about spreadsheet errors and in particular about the classification of those errors is already going on for decades. Panko and Halverson Jr. (1996) distinguishes research on spreadsheet errors and risks along certain dimensions, namely life cycle stage, research issue, and methodology. However, since the present work is not about spreadsheet errors as such, but about shortcomings of spreadsheets with respect to the support of business processes, it cannot be classified by this classification scheme. Furthermore, Panko and Halverson Jr. (2000) propose various dimensions for differentiating errors, e.g., qualitative vs. quantitative errors. Based on their work, Rajalingham et al. (2008) designed a framework for systematically classifying spreadsheet errors based on the nature and characteristics of those errors, which consists of a taxonomy of spreadsheet errors. Around the research on spreadsheet errors, many related research activities evolved focusing on the prevention and elimination of them (Powell et al., 2008). For example, Abraham and Erwig (2005) propose an approach for debugging spreadsheets. On the other hand, in order to already prevent spreadsheet errors during the spreadsheet's design, Cunha et al. (2012), Engels and Erwig (2005), and Janvrin and Morrison (2000) describe structured spreadsheet design approaches and model-based engineering approaches respectively.

Another related research stream is about the role of the user in the design and usage of spreadsheets. Scaffidi, Shaw, et al. (2005) anticipated that in 2012 over 55 million end-users (Nardi, 1993) in the US are using spreadsheets. However, as shown by Brown and Gould (1987) and Gibbs et al. (2014), there is still a huge variety of skill levels among those users regarding the design and usage of spreadsheets, although the knowledge about spreadsheet design and usage has a direct impact onto the success of the application of spreadsheets (McGill and Klobas, 2005). Scaffidi, Ko, et al. (2006) deal with the question of how end-users are actually using spreadsheets. More specifically, Lawson et al. (2009) and Pemberton and Robson (2000) study which spreadsheet functions end-users are actually using in their daily work. Thus, they study the performance of users when designing and using spreadsheets, while the present work focuses on the performance of spreadsheets regarding the support of business processes.

The relevance of spreadsheets for the business was highlighted by Grossman et al. (2005). They claim that spreadsheets function as information systems, and that there is a heterogeneity of purposes for using spreadsheets. Thereby they emphasize the need for a taxonomy of spreadsheet information systems, which coincides with the purpose of the present paper. In another work, Chan and Storey (1996) investigate the correlation between spreadsheet proficiency and business tasks users have to perform. Therefore they also relate the spreadsheet to its actual business context. Furthermore, Panko (2006), Pemberton and Robson (2000), and Winston (2001) describe different application areas of spreadsheets, e.g., financial reporting, workload planning, and quality control, and how spreadsheets are basically applied in these areas. By doing this, Pemberton and Robson (2000) are already characterizing spreadsheets along usage-related dimensions (e.g., data transfer). However, on the one hand they primarily focus on functional characteristics of spreadsheets, while the present paper also considers non-functional aspects (e.g., *Dimension 1: Design Context* as described in Section 4.1). On the other hand, they do not identify clusters or patterns of spreadsheet usages. Nardi and Miller (1990) and Panko and Port (2012) claim that concerns like collaboration, privacy, security, and compliance are essential for today's companies. Although they consider spreadsheets as information systems in an organizational context, and furthermore underline the existence of those concerns, they do not differentiate between types or patterns of spreadsheet usages, and thus do not examine if those concerns only occur in certain types or patterns of spreadsheet usages. The present paper might be a step towards such an differentiation. Ronen et al. (1989) propose the concept of design context for spreadsheets which captures the spreadsheet's usage scope and frequency. As later on described in Section 4.1, we define the design context indeed as one classification dimension. Furthermore, Ronen et al. (1989) use a data flow representation of spreadsheets as a structured way of designing spreadsheets. In the present paper, we use a similar notion for describing spreadsheet usages and also for deriving classification dimensions in order to be able to identify spreadsheet usage patterns. However, in the present work we added further details to the information flow diagrams, e.g., the spreadsheet's data origin and its data consumers.

3 Case Studies

In this section, we describe the setting and results from case studies about the usage of spreadsheets which we conducted in 2 German companies. The general research methodology was exploratory case research as described by Benbasat et al. (1987). In particular, we applied the analytic technique of explanation building (Yin, 2014) for multiple-case studies. The goal of the studies was to discover how today's companies are using spreadsheets, and what are typical shortcomings of spreadsheets in these usage scenarios. The identified use cases form the foundation for the development of a framework for classifying spreadsheet usages. Thereby, we cooperated with two companies—a big logistics and courier company as well as a financial risk management research laboratory. In the remainder of this paper we refer to those companies with "Company 1" and "Company 2" respectively.

3.1 Case Study Procedures

The case research was organized as follows:

1. As a first step, we asked the two companies for critical spreadsheets they are using in their daily work, and where they already face shortcomings when using spreadsheets to support respective business processes. Company 1 responded with four spreadsheet usages, while Company 2 responded with five, so that in total we observed nine spreadsheet usages as well as their respective contexts.

2. In a second step, we arranged and conducted interviews with the responsible users and designers of the given spreadsheets. The interviews covered the following interview sections:

   Life-cycle and design context This interview section was about the spreadsheet's creation date, the rate of change of its design and data, and the number of instances of the spreadsheet.

   User-related and collaborative aspects In this interview section, we asked about the users of the spreadsheet as well as their skills, responsibilities, and access rights. Furthermore, we were interested in how spreadsheets are transferred from one workplace to another, how users know when to interact with the spreadsheet, and how new users are familiarized with the spreadsheet.

   Data-related aspects Data-related questions were about the origin of the spreadsheet's data, how and which type of data is entered, which kind of calculations and transformations take place in the spreadsheet, and who or what are the actual consumers of the spreadsheet's output data.

   Usage-related aspects This interview section was about the shortcomings faced when using the spreadsheets for a certain purpose, and why spreadsheets are used despite having these shortcomings. Furthermore, we were interested if other tools are considered for replacing the spreadsheet in the respective use case. The information gathered from this section is particularly relevant for future research about shortcomings of certain spreadsheet usage patterns.

3. In addition to conducting the interviews, we also asked for the (anonymous) spreadsheets themselves as well as respective documentations in order to have additional sources of evidence. While we got all the spreadsheets from Company 1, the algorithms in the spreadsheets of Company 2 constitute trade secrets, wherefore we got screen-shots illustrating the basic structure and content of their spreadsheets.

4. As a next step, we consolidated the information gathered from the interviews, the spreadsheets, and respective documentations and translated this into a unified and consistent form. As proposed by Eisenhardt and Graebner (2007), we are using a visual notation for describing the cases.

5. Thereafter, the consolidated information was validated again by the responsible persons of the respective spreadsheet applications, which corresponds to the application of the hermeneutic circle (Cole and Avison, 2007; Myers, 2004).

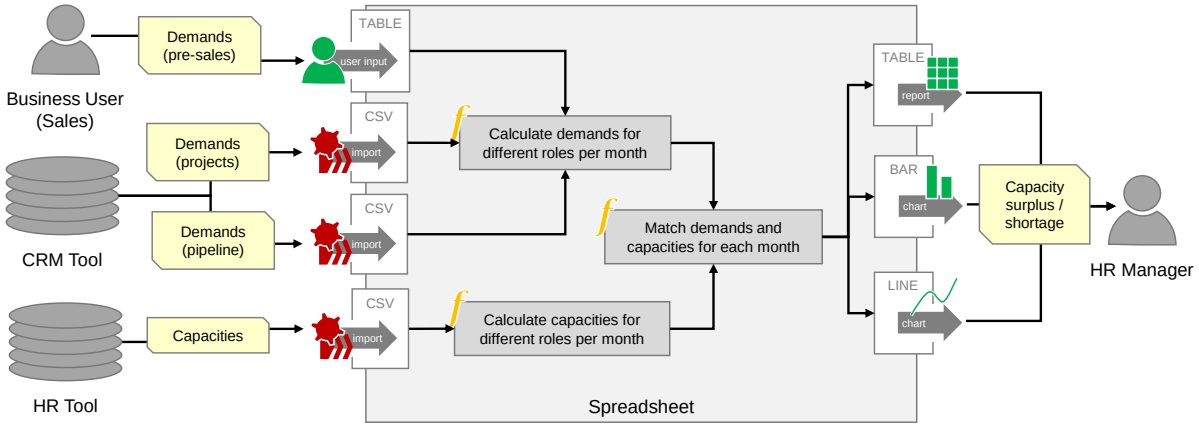

Figure 2: Information flow representation of Case 1.A: Capacity Planning.

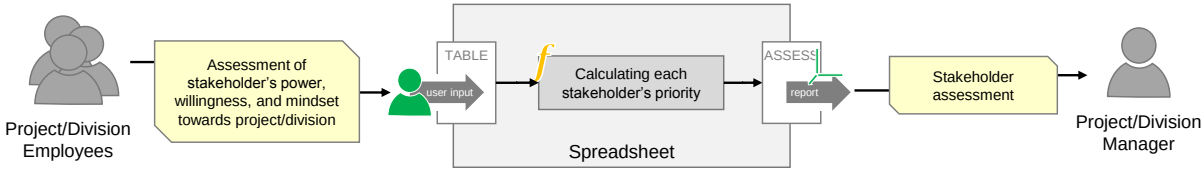

Figure 3: Information flow representation of Case 1.C: Stakeholder Analysis.

3.2 Documentation of Observed Cases

The results of the conducted case studies are documentations of nine usages of spreadsheets supporting various kinds of business processes. The observed use cases of Company 1 are as follows:

1.A: Capacity Planning The purpose of this spreadsheet is the comparison of Human Resources (HR) with demands induced by current projects, pipelined projects, and pre-sales/3rd-level support activities. Thereby, the spreadsheet supports workforce managers in identifying shortages and surpluses of HRs in order to plan future HR activities. The results of the spreadsheet are charts visualizing demands vs. capacities as well as tables with more detailed information about demands and capacities per month.

1.B: IT Financial Reporting The spreadsheet maps IT cost records imported from a financial reporting system to respective IT divisions in order to determine the distribution of all IT costs over divisions. The resulting charts are exported to a presentation software for reporting purposes.

1.C: Stakeholder Analysis For each IT project, the spreadsheet contains a list of all stakeholders. Moreover, the stakeholders are assessed and categorized with respect to their power within the organization, willingness to engage, and mindset towards the respective project and visualized in a three-dimensional matrix. The spreadsheet is maintained by multiple members of the project team.

1.D: Risk Management For each IT project, the spreadsheet contains a list of identified risks and proper mitigation and contingency actions. Moreover, the risks are assessed and prioritized with respect to their probability and impact and visualized in a two-dimensional matrix. Again, the spreadsheet is maintained by multiple members of the project team.

The investigated use cases of Company 2—a financial risk management research laboratory—are:

2.A: Performance Calculation The spreadsheet imports financial data from a relational database and calculates returns of investments (ROI) based on certain parameters provided by an investment advisor. The output of the spreadsheet is imported by another software system and used for the determination of future investment strategies.

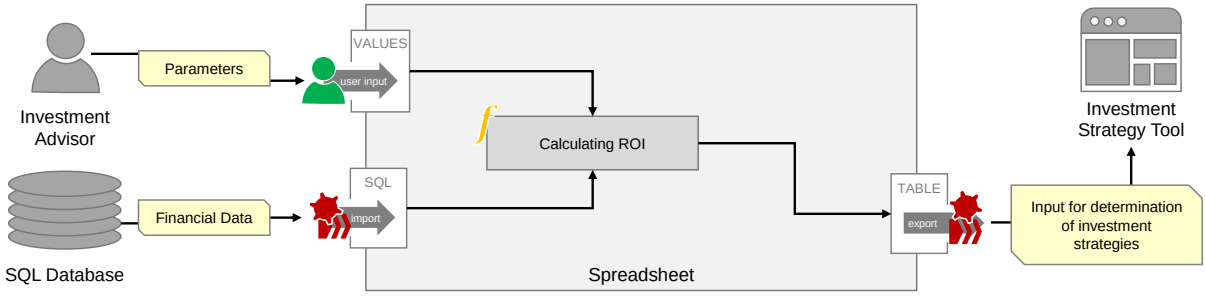

Figure 4: Information flow representation of Case 2.A: Performance Calculation.

2.B: Simulation Result Reporting The spreadsheet receives data and parameters from a simulation software generating different scenarios for a given time-span. The spreadsheet generates a two-fold output: On the one hand, it generates graphs which are exported to a presentation software for direct reporting purposes. On the other hand, the spreadsheet transforms the data to a format which can be imported by another software system.

2.C: Liability Index Calculation This spreadsheet calculates and discounts cash-flows based on parameters (e.g., current interests rates) provided by an asset liability strategy manager. The purpose of this spreadsheet is the validation against the results determined by another software system.

2.D: Time-series Transformation This spreadsheet enriches financial data gathered from a relational database with information extracted from a report from a financial services provider. Thereby, the spreadsheet generates time-series of cash-flows and ROIs, which in turn are exported to a database.

2.E: Scenario Cube Validation By importing pre-processed statistics from multiple scenarios computed by a numerical computing software, the spreadsheet transforms the data in a way so that it is able to generate a visual report consisting of multiple charts of different types.

As proposed by Eisenhardt and Graebner (2007), we created a visual documentation of the spreadsheet's usage for each of those nine cases (Figures 2, 3, and 4 illustrate the spreadsheets of cases 1.A, 1.C, and 2.A respectively). These visual documentations depict the information flow from the original data sources of spreadsheets through to the consumers of their output. Ronen et al. (1989) already used flow diagrams for describing spreadsheets and how they are used. However, while Ronen et al. (1989) capture the spreadsheet's input and output as well as the computations within the spreadsheet, they neglect the origin and consumers of the spreadsheet's information on the one hand, and how the spreadsheet's input and output are entered and represented respectively on the other hand.

The visual documentation representing the abstract information flow of a certain spreadsheet usage captures the type of data origin, the data input method, the abstract data transformation, the type of the output view, as well as the type of data consumer. Thereby, it explicitly differentiates, e.g., between manual and automatic input, or between human or artificial consumers of the spreadsheet's data. By this visual notation, the spreadsheet usages can be represented in a unified and consistent way.

4 A Framework for Classifying Spreadsheet Usages

In the context of the present work, a spreadsheet usage pattern represents a class of similar spreadsheet usages with respect to certain aspects. Thereby, based on a cross-case view and in particular based on the information flow representations of spreadsheet usages as described in the previous section, and as proposed by Eisenhardt (1989), we identified seven dimensions capturing usage-related aspects of spreadsheets. Figure 5 illustrates how the dimensions were derived. Furthermore, for each of those dimensions we derived a small set of possible categories, wherefore the framework forms a morphological box as shown in Figure 6. The morphological box as defined by Zwicky (1969) is a multi-dimensional

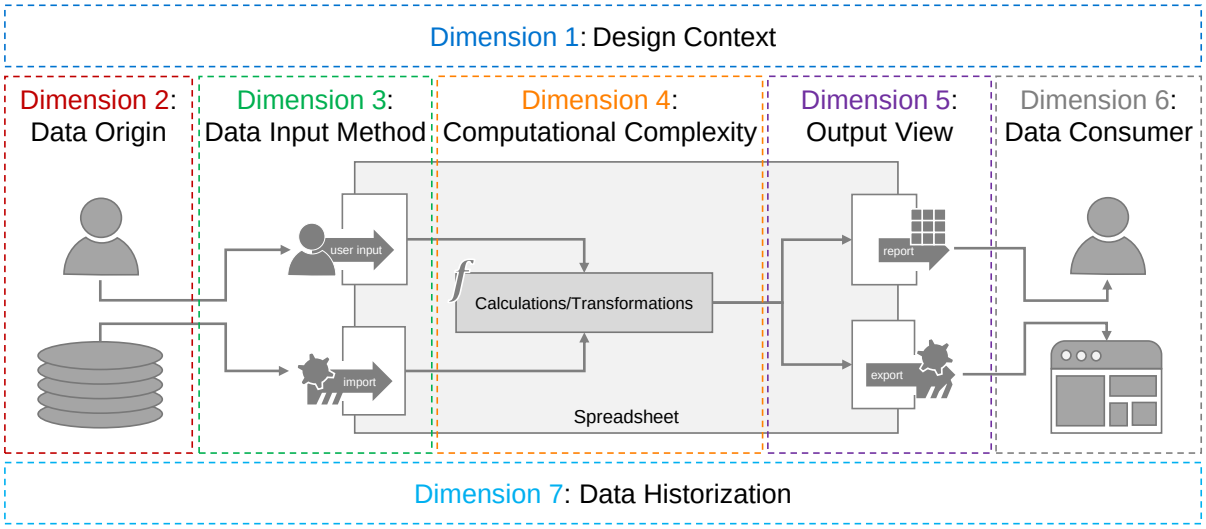

Figure 5: Identifying spreadsheet usage classification dimensions based on the information flow representation.

Design Context	Data Origin	Data Input Method	Computational Complexity	Output View	Data Consumer	Data Historization
Frequent usage by multiple users	User Assessment	Manual Input	Low Complexity	Raw Data	Software Systems	Accumulation of Data
Frequent usage by single user	Unstructured Data Source	Hybrid Input	Medium Complexity			Versioning of File
One-shot	Structured Data Source	Automatic Import	High Complexity	Dashboard	Humans	No Historization

Figure 6: The derived classification framework as a morphological box.

matrix and a tool for the morphological analysis, whereas this analysis method captures the concise formulation of the respective problem, the identification of all parameters/dimensions which might be relevant for the problem, and the subsequent derivation of possible solutions for each of those parameters. The following subsections describe the design of the classification framework in detail.

4.1 Dimension 1: Design Context

As defined by Ronen et al. (1989), a spreadsheet's design context captures if the spreadsheet is used just by its designer or also by other users, and how frequently the spreadsheet is actually used. The design context of a spreadsheet influences the way it should be designed. For example, if a spreadsheet targets multiple users, and/or has to be used very frequently, it should be subject of formal design procedures, while in case of a spreadsheet which is only used once (*one-shot*) there is no need for an extensive design process. For example, the spreadsheets in cases 1.C and 1.D are both designed for multiple usages by multiple users, while the spreadsheets of cases 2.A and 2.D are just used by the creator himself. Although there is no one-shot spreadsheet among the observed cases (since we asked the companies for critical spreadsheets they are using in their daily work), the work by Ronen et al. (1989) suggests that the design context dimension should also capture this class of spreadsheets.
Hence, we derived the following set of categories for the classification dimension design context:

Frequent usage by multiple users A spreadsheet of this category is designed for a long-term usage and used multiple times by the creator as well as other users. Usually, the creator has the role of a data

modeler and is responsible for the spreadsheet's design and overall structure, while other users are primarily data providers.

Frequent usage by single user A spreadsheet of this category is designed for a long-term usage, but only used by the creator himself. Usually, the purpose of this kind of spreadsheets is the support of recurring data-based tasks and activities.

One-shot A spreadsheet of this category is created ad-hoc for one single purpose. After using the spreadsheet for this purpose, the spreadsheet is usually disposed.

4.2 Dimension 2: Data Origin

As illustrated in Figure 5, the data origin dimension of the classification framework captures the provenance of the data (Simmhan et al., 2005), i.e., the type of the data source which provides the input data for the spreadsheet. Thereby, we differentiate between objective facts as determined by technical data sources on the one hand, and (potentially subjective) assessments made by humans on the other hand. Furthermore, this dimension also deals with the structuredness of the data sources. For example, the input of the spreadsheets in cases 1.C and 1.D is solely consisting of information assessed by a group of users. In contrast, the spreadsheet in case 2.D is manually enriched with information extracted from a financial report, while in case 2.A the data is directly imported from a SQL database.

Therefore, we derived the categories for the classification dimension data origin as follows:

User Assessment Most of the input of spreadsheets of this category has no concrete technical data source, i.e., the spreadsheet's input primarily consists of data assessed by a user or a user group. These spreadsheets are usually used for supporting cognitive tasks, e.g., the assessment and evaluation of risks.

Unstructured Data Source The origin of most of the spreadsheet's input data is unstructured data like PDF reports, emails, and charts. Therefore, the purpose of entering this input data into the spreadsheet is usually the manual structuring and extraction of information.

Structured Data Source Most of the spreadsheet's data comes from structured data sources (e.g., relational databases, CSV, web services) whose data is usually automatically importable through an API (application programming interface).

4.3 Dimension 3: Data Input Method

The third classification dimension deals with the method of the data input, i.e., how the spreadsheet's input data is actually entered (c.f. Figure 5). Thereby, we differentiate between manual input, automatic import, and a hybrid approach. For example, cases 1.C and 1.D are solely based on manual input, while the data for the spreadsheet of case 2.B is solely automatically imported from an SQL database. In case 1.A, a user enters some data manually in addition to the automatically imported data. In cases 2.B and 2.C, the calculation and transformation of the automatically imported data is configured through parameters manually entered by the user of the spreadsheet.

Thereby, we distinguish between the following data input methods:

Manual Input The input data of spreadsheets of this category has to be entered manually by one or multiple users.

Hybrid Input The majority of the input data of this category's spreadsheets is imported automatically. However, the spreadsheet also requires manual input by users, e.g., to enrich the automatically imported data with further information, or to configure the calculation and transformation procedures within the spreadsheet by providing parameters.

Automatic Import The entire input data of spreadsheets of this category is imported automatically from another software system, e.g., from a SQL database or CSV file. The spreadsheet does not require any further manual interaction with its input data.

4.4 Dimension 4: Computational Complexity

As depicted in Figure 5, the fourth dimension of the classification framework captures the computational complexity of the spreadsheet. Thereby spreadsheets are classified with respect to the complexity of formulas and operations they are using for transforming the input data into the output data. Hall (1996) also captured the complexity of spreadsheets by differentiating between layout, logical, and link complexity, whereas the logical one refers to the computational complexity as depicted in Figure 5. For example, while the spreadsheets of cases 1.C and 2.A are both using just elementary arithmetic operations and/or conditionals, the spreadsheet of case 1.A is using more complex operations like matrix multiplications. The spreadsheet of case 2.E is not even implementable solely with the default operations of the spreadsheet system, i.e., it requires certain extensions to spreadsheets.
Therefore, we derived the following categories for the computational complexity dimension:

Low Complexity Spreadsheets of this category are only using elementary operations, e.g., simple arithmetic operations (e.g., addition, subtraction), aggregation operations of data rows and columns (e.g., sum, average), and simple conditionals.

Medium Complexity Spreadsheets of this category are using advanced spreadsheet functions, which usually take multidimensional cell ranges as an input for performing, e.g., matrix multiplications or advanced statistical methods.

High Complexity In this category, the default computation capabilities of prevalent spreadsheet tools are not sufficient for supporting the respective business process. Therefore, the spreadsheet application has to be extended by additional functionality (e.g., by macros in MS Excel).

4.5 Dimension 5: Output View

The fifth dimension of the classification framework (c.f. Figure 5) is the type of the output of the spreadsheet, i.e., if the output consists of interactive or visual elements. For example, the spreadsheets of cases 1.A and 2.E are producing dashboards consisting of multiple charts and pivot tables. In contrast, the spreadsheets of cases 2.C and 2.D are just producing flat and unformatted data tables.
Hence, the categories of the output view dimension are derived as follows:

Raw Data This category's spreadsheets are primarily producing one or multiple simple and unformatted data tables. These outputs are neither interactive nor do they contain visual elements, e.g., color-encoded information.

Dashboard The primary output of spreadsheets of this category is one or multiple dashboards, which in turn consist of one or multiple visualizations or interactive pivot tables.

4.6 Dimension 6: Data Consumer

The sixth dimension of the classification framework is the type of the consumer of the spreadsheet's output. Thereby we differentiate between human and artificial consumers. For example, the output of the spreadsheets of cases 1.A and 1.C is consumed by users (e.g., by a HR manager and project manager respectively) for decision support purposes. In contrast, the output of the spreadsheets of cases 1.B and 2.A is the input for another software system, e.g., a presentation software or data base system.
Based on this observation, we derived the following categories for the data consumer dimension:

Software Systems The primary consumers of the output of this category's spreadsheets are software systems. For example, such a spreadsheet might generate a data table, which is imported to a database system, or a chart, which is embedded in a presentation software.

Humans The primary consumers of the output of this category's spreadsheets are human users. Usually, the purpose of this kind of spreadsheets is the visual processing of the spreadsheet's data.

4.7 Dimension 7: Data Historization

The last dimension captures how the spreadsheet's data is historized, i.e., if and how previous versions of the spreadsheet's data are managed by the spreadsheet user. As a study of Lawson et al. (2009) has shown, tracking the evolution of the spreadsheet's data over time is very common in practice. For example, in cases 1.A and 2.D new data is only appended to the existing one, i.e., the spreadsheet keeps track of the evolution of the data. Another way of historizing the spreadsheets data is to make regular snapshots of the whole spreadsheet (e.g., as it is done in cases 1.C and 2.A). Furthermore, Lawson et al. (2009) revealed that also the application of version control systems is a proper way of historizing a spreadsheet's data. This data historization strategy is also applied in case 2.E.
Therefore, we derived the following categories for the data historization dimension:

Accumulation of Data In spreadsheets of this category, new data is appended to existing one, i.e., existing data remains in the spreadsheet. Thereby, the spreadsheet explicitly models its data history and thus enables also the analysis of the evolution of the spreadsheet's data.

Versioning of File Users of this category's spreadsheets are either making snapshots of the current spreadsheet in order to store its current state, or they use a version control system to do versioning of the whole spreadsheet file. New file versions are either made on a regular basis (e.g., once a month) or always right before the change of the spreadsheet's data. Since the spreadsheet itself does not capture the evolution of the data, the temporal aspect of the data cannot be subject of data analysis.

No Historization Although in all cases we have observed at least one of the aforementioned data historization approaches, we are also considering the category of spreadsheets, which are neglecting the temporal evolution of their data. This means that in these spreadsheets, the already existing data is replaced by the new one, without doing any versioning of the spreadsheet.

5 Identification of Usage Patterns

In order to identify spreadsheet usage patterns, i.e., spreadsheet usages which are similar with respect to the aspects captured by the classification framework's dimensions, we applied a clustering algorithm to the classifications of all cases. Thereby, we encoded the classifications of all cases in a matrix as shown in Table 1. In this matrix, the rows represent the use cases, and the columns the dimensions of the morphological box. Since there is an implicit order of classes within each dimension, the dimensions' scales are ordinal, and we can map the ordinal value to a corresponding numerical value. The arithmetic difference between two numerical values represents the diversity between two ordinal values. For example, a spreadsheet with design context *Frequent usage by multiple users* is intuitively more diverse from a spreadsheet with design context *One-shot* than from a spreadsheet with design context *Frequent usage by single user*. Hence, in case of ternary dimensions (e.g., design context), we map each ordinal value to 0, 2, or 4. By assuming that the difference between two values of a binary dimension has to be greater than between two adjacent values in a ternary dimension, but less than between two non-adjacent values, we map the ordinal values of the binary dimensions to 0 and 3 respectively.
By the aforementioned encoding of the classifications of each of the nine cases we were able to apply the k-means clustering algorithm. Thereby, based on a classification matrix X, the total distance between two spreadsheets i and j can be determined by the Euclidean distance in a seven-dimensional space defined as $\sqrt{\sum_{d=1}^{7}(X_{i,d}-X_{j,d})^2}$. The optimal number of clusters for the classification matrix in Table 1 was determined by the elbow criterion. Thereby, we plotted the variance of the clusters against the number of clusters (c.f., Figure 7), whereas the elbow in the plot determines three as the optimal number of clusters. Therefore, the application of the k-means algorithm to the nine spreadsheet usages as described in Section 3 results in three clusters representing corresponding spreadsheet usage patterns, which we name and describe as follows:

	D1	D2	D3	D4	D5	D6	D7
1.A	2	4	2	2	3	3	0
1.B	2	4	4	0	3	0	2
1.C	0	0	0	0	3	3	2
1.D	0	0	0	0	3	3	2
2.A	2	4	2	0	0	0	2
2.B	2	4	4	0	0	0	2
2.C	2	4	2	0	0	3	2
2.D	2	4	2	2	0	0	0
2.E	2	4	2	4	3	3	0

Table 1: The classification matrix, whereas rows represent the nine cases of Section 3, and columns represent the seven dimensions of the morphological box (c.f., Figure 6)

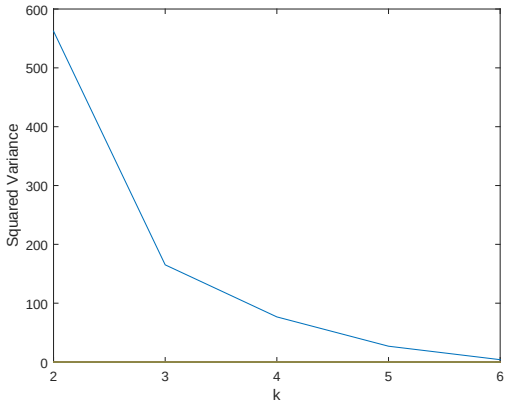

Figure 7: Application of the elbow criterion to determine the number of clusters

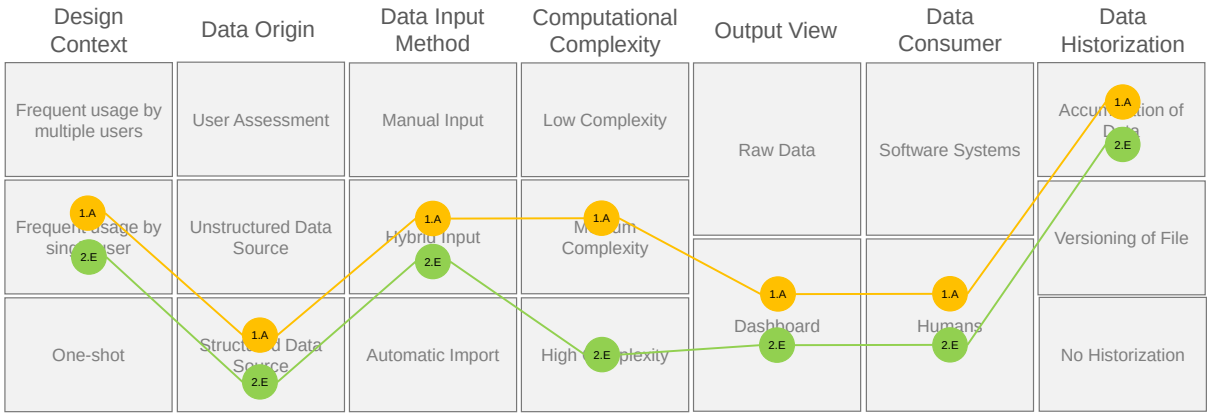

Figure 8: Spreadsheet usage pattern *Reporting Sheet*

Reporting Sheets The first pattern is the *Reporting Sheet* and includes the spreadsheets of cases 1.A and 2.E (c.f., Figure 8). *Reporting Sheets* usually integrate and consolidate data from various and multiple data sources, apply rather complex data transformations and analysis, and visualize the results in form of a dashboard consisting of potentially multiple charts and tables (c.f., Figure 2).

Documentation Sheets The second cluster represents the pattern of *Documentation Sheets* consisting of the spreadsheets of cases 1.C and 1.D (c.f., Figure 9). *Documentation Sheets* usually collect information from multiple users, and apply neither complex data transformations nor do they generate complex views on the documented data (c.f., Figure 3).

Data Transformation Sheets The last pattern identified based on the nine cases is the *Data Transformation Sheet* consisting of the spreadsheets of cases 1.B, 2.A, 2.B, 2.C, and 2.D (c.f., Figure 10). Those sheets usually take the data from one input source (eventually enriched by manually entered data) and apply rather simple data transformations. Usually the resulting output of those sheets is provided in a way so that it can serve as an input for another software system (c.f., Figure 4).

6 Threats to Validity

Perry et al. (2000) distinguish between three kinds of validity, namely construct validity, internal validity, and external validity.

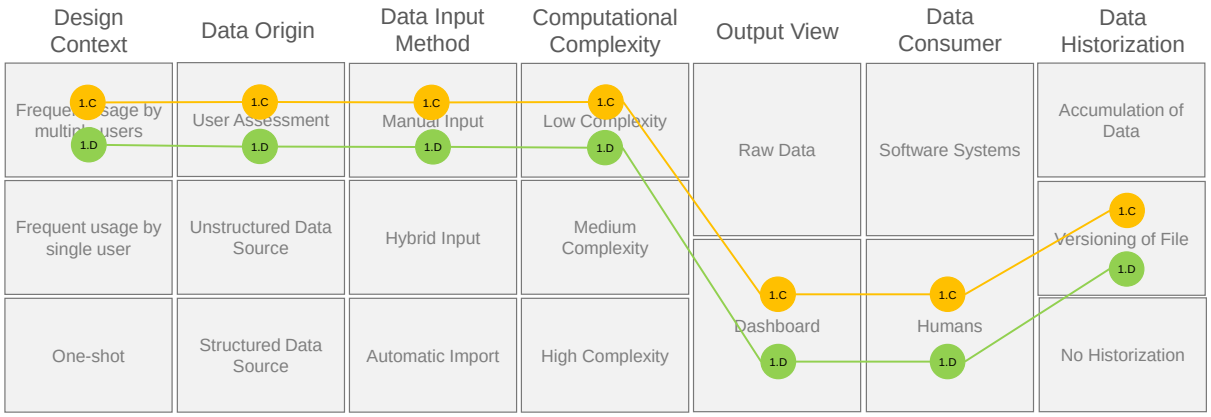

Figure 9: Spreadsheet usage pattern *Documentation Sheet*

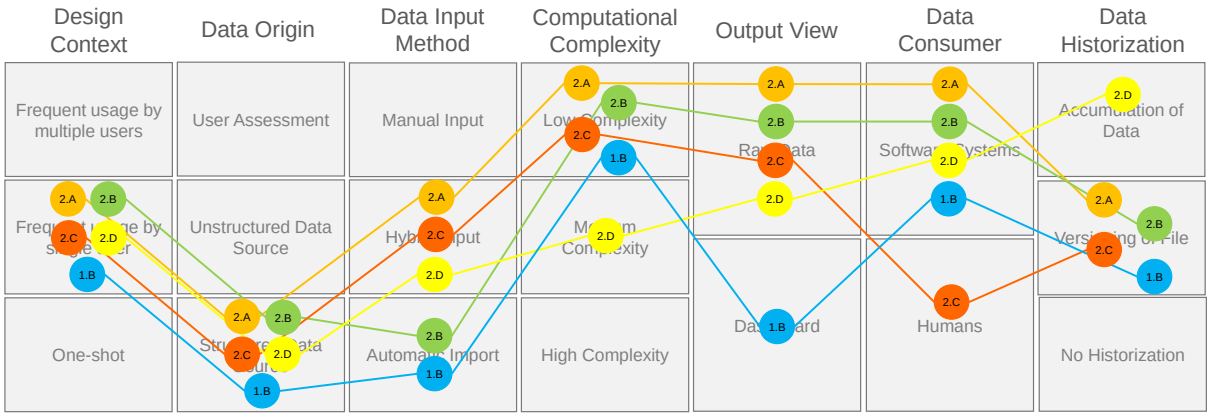

Figure 10: Spreadsheet usage pattern *Data Transformation Sheet*

Construct validity refers to the appropriateness of the case study design and analysis approach with respect to the research objective. As stated in Section 1, our research objective is not to study the syntax or semantics of spreadsheets, but how they are applied for supporting certain business processes. Therefore, in our studies we considered spreadsheets as black-boxes (although the computational complexity dimension indicates the spreadsheet's internal structure), as we were more interested in their context, and how users interact with spreadsheets. Hence, we explicitly captured usage-aspects of spreadsheets rather than syntactic or semantic properties. Furthermore, developing a classification framework consisting of multiple dimension as described in Section 4 is a common approach for identifying cross-case patterns (Dubé and Paré, 2003; Eisenhardt, 1989), wherefore we consider the applied research approach appropriate for the targeted research objective.

With respect to the internal validity, we admit that nine cases of spreadsheet usages form a small foundation for the identification of patterns through clustering. This is due to the fact that studying not only the spreadsheets, but also and in particular their context and actual usage entails a relatively high effort including the preparation, conduction, and analysis of interviews as well as the analysis of the spreadsheets and their documentations. Nevertheless, based on those nine cases we derived a classification framework which can serve as a foundation for a well-structured survey and thus enables a quantitative study on spreadsheet usage patterns. This has to be done in a new study, though. Another threat to the internal validity of the present work is the derivation of the classification framework. While the derivation of each dimension of the morphological box is described in Section 4 and based on the observed use cases as well as related literature, there are certainly many other aspects of spreadsheet usages which are not

captured by the morphological box. However, we believe that the identified set of dimensions and their classes is appropriate with respect to its purpose of deriving spreadsheet usage patterns. Nevertheless, future research activities building upon the proposed solution can still extend the morphological box by additional dimensions and classes. Apart from the derivation of the classification framework, a discussion of the resulting spreadsheet usage patterns with experts from the two companies revealed that the identified patterns match their personal perception of types of spreadsheet usages.

The external validity of a study determines the generalizability of the results to settings outside the study. While the derived classification framework can be applied to any other spreadsheet usages, the identification of spreadsheet usage pattern strongly depends on the actual set of observed cases. This means that the application of the clustering algorithm on a different set of spreadsheets might yield to different spreadsheet usage pattern. However, on the one hand we believe that those will be similar to or refinements of the patterns as described in Section 5. And on the other hand, applying the classification framework to spreadsheets of a company and subsequently clustering them to identify company-specific patterns still supports the respective company in evaluating the suitability of a spreadsheet cluster for the support of a certain business process. Furthermore, in future we want to conduct a quantitative study of spreadsheet usages based on the proposed classification framework, which could either confirm the validity of the usage patterns as derived in Section 5, or extend the set of patterns by further ones.

7 Conclusion and Future Work

In this paper, we have presented the results of an empirical study that we conducted in two companies in order to identify spreadsheet usage patterns by a cross-case analysis. Thereby, we proposed a morphological box for the classification of spreadsheets along seven usage-related dimensions (c.f., Section 4), which is the answer to the first research question raised in Section 1. The second research question was answered in Section 5, which describes the determination of spreadsheet usage patterns based on the clustering of classified spreadsheets. In Section 6 we discussed threats to the validity of the study and already proposed further studies in this area, in particular a quantitative study for strengthening the validity of the identified patterns and determination of further ones.

Based on the results of our work, future research activities can focus on shortcomings of spreadsheet usages with respect to their support of business processes, e.g., data-row-based access control, managed evolution of multiple spreadsheet instances, and support for complex data-structures. Therefore, in our case studies we not only asked for the context of the respective spreadsheets, but also for this kind of shortcomings. While currently those shortcomings are associated with the specific cases, spreadsheet usage patterns as identified in this work enable the detachment of the shortcomings from the concrete cases and the linkage of them to the more abstract patterns. Hence, future research can address specific spreadsheet usage patterns by further studying certain patterns or by designing tailored solution approaches and tools for particular spreadsheet usage patterns. For example, as mentioned in Section 2, the findings by Nardi and Miller (1990) as well as Panko and Port (2012) regarding concerns of spreadsheets can be revisited in the light of spreadsheet usage patterns as identified in the present work. Furthermore, based on the identified usage patterns as well as pattern-specific concerns, researchers are able to design pattern-specific solutions on the one hand, and practitioners are able to reevaluate the application of spreadsheets for usages of a certain pattern on the other hand, e.g., by establishing organization-specific guidelines for documentation spreadsheets. Therefore, the present paper is valuable for both research and practice.

Acknowledgement

This research has been sponsored in part by the German Federal Ministry of Education and Research (BMBF) with grant number TUM: 01IS12057.

References

Abraham, R. and M. Erwig (2005). "Goal-Directed Debugging of Spreadsheets." *Proceedings of the Symposium on Visual Languages and Human-Centric Computing*, 37–44.

Basel II (2004). Basel Committee on Banking Supervision. URL: http://www.federalreserve.gov/boarddocs/press/bcreg/2004/20040626/attachment.pdf.

Benbasat, I. et al. (1987). "The Case Research Strategy in Studies of Information Systems." *Management Information Systems Quarterly*, 369–386.

Braa, K. and R. Vidgen (1999). "Interpretation, Intervention, and Reduction in the Organizational Laboratory: A Framework for In-Context Information System Research." *Accounting, Management and Information Technologies* 9 (1), 25–47.

Bradley, L. and K. McDaid (2009). "Using Bayesian Statistical Methods to Determine the Level of Error in Large Spreadsheets." *Proceedings of the International Conference on Software Engineering*, 351–354.

Brown, P. S. and J. D. Gould (1987). "An Experimental Study of People Creating Spreadsheets." *ACM Transactions on Information Systems* 5 (3), 258–272.

Caulkins, J. P. et al. (2007). "Spreadsheet Errors and Decision Making: Evidence from Field Interviews." *Journal of Organizational and End User Computing* 19 (3), 1–23.

Chan, Y. E. and V. C. Storey (1996). "The Use of Spreadsheets in Organizations: Determinants and Consequences." *Information & Management* 31 (3), 119–134.

Cole, M. and D. Avison (2007). "The Potential of Hermeneutics in Information Systems Research." *European Journal of Information Systems* 16 (6), 820–833.

Cunha, J. et al. (2012). "MDSheet: A Framework for Model-driven Spreadsheet Engineering." *Proceedings of the International Conference on Software Engineering*, 1395–1398.

Davis, G. B. and M. H. Olson (1985). *Management Information Systems: Conceptual Foundations, Structure, and Development*. 2nd Edition. McGraw-Hill, Inc. ISBN: 0-07-015828-2.

Dubé, L. and G. Paré (2003). "Rigor in Information Systems Positivist Case Research: Current Practices, Trends, and Recommendations." *Management Information Systems Quarterly*, 597–636.

Eisenhardt, K. M. (1989). "Building Theories from Case Study Research." *Academy of Management Review* 14 (4), 532–550.

Eisenhardt, K. M. and M. E. Graebner (2007). "Theory Building from Cases: Opportunities and Challenges." *Academy of Management Journal* 50 (1), 25–32.

Engels, G. and M. Erwig (2005). "ClassSheets: Automatic Generation of Spreadsheet Applications from Object-Oriented Specifications." *Proceedings of the International Conference on Automated Software Engineering*, 124–133.

Gibbs, S. et al. (2014). "Are Workplace End-User Computing Skills at a Desirable Level? A New Zealand Perspective." *Proceedings of the Americas Conference on Information Systems*.

Grossman, T. A. et al. (2005). *Spreadsheet Information Systems are Essential to Business: working paper*.

Hall, M. J. J. (1996). "A Risk and Control-Oriented Study of the Practices of Spreadsheet Application Developers." *Proceedings of the Hawaii International Conference on System Sciences*, 364–373.

Hermans, F. (2012). "Analyzing and Visualizing Spreadsheets." PhD thesis. Technische Universiteit Delft.

Janvrin, D. and J. Morrison (2000). "Using a Structured Design Approach to Reduce Risks in End User Spreadsheet Development." *Information & Management* 37 (1), 1–12.

Lawson, B. R. et al. (2009). "A Comparison of Spreadsheet Users with Different Levels of Experience." *Omega* 37 (3), 579–590.

McGill, T. J. and J. E. Klobas (2005). "The Role of Spreadsheet Knowledge in User-Developed Application Success." *Decision Support Systems* 39 (3), 355–369.

MetricStream GRC Blog (2014). *The True Cost of Spreadsheets in Risk and Compliance Management*. Ed. by MetricStream GRC Blog. URL: http://blog.metricstream.com/2014/cost-of-spreadsheets-in-risk-and-compliance-management/ (visited on 12/17/2014).

Myers, M. D. (2004). "Hermeneutics in Information Systems Research." *Social Theory and Philosophy for Information Systems*, 103–128.

Nardi, B. A. (1993). *A Small Matter of Programming: Perspectives on End User Computing*. MIT Press. ISBN: 0262140535.

Nardi, B. A. and J. R. Miller (1990). "An Ethnographic Study of Distributed Problem Solving in Spreadsheet Development." *Proceedings of the Conference on Computer-Supported Cooperative Work*, 197–208.

Panko, R. R. (2006). "Facing the Problem of Spreadsheet Errors." *Decision Line* 37 (5).

— (2008). "Spreadsheet errors: What we know. what we think we can do." *arXiv preprint arXiv:0802.3457*.

Panko, R. R. and R. Halverson Jr. (1996). "Spreadsheets on Trial: A Survey of Research on Spreadsheet Risks." *Proceedings of the Hawaii International Conference on System Sciences*, 326–335.

— (2000). "Two Corpuses of Spreadsheet Errors." *Proceedings of the Hawaii International Conference on System Sciences*, 1–8.

Panko, R. R. and D. N. Port (2012). "End User Computing: The Dark Matter (and Dark Energy) of Corporate IT." *Journal of Organizational and End User Computing*, 4603–4612.

Pemberton, J. D. and A. J. Robson (2000). "Spreadsheets in Business." *Industrial Management & Data Systems* 100 (8), 379–388.

Perry, D. E. et al. (2000). "Empirical Studies of Software Engineering: A Roadmap." *Proceedings of the Conference on the Future of Software Engineering*, 345–355.

Powell, S. G. et al. (2008). "A Critical Review of the Literature on Spreadsheet Errors." *Decision Support Systems* 46 (1), 128–138.

— (2009). "Impact of Errors in Operational Spreadsheets." *Decision Support Systems* 47 (2), 126–132.

Rajalingham, K. et al. (2008). "Classification of Spreadsheet Errors." *arXiv preprint arXiv:0805.4224*.

Ronen, B. et al. (1989). "Spreadsheet Analysis and Design." *Communications of the ACM* 32 (1), 84–93.

Sarbanes-Oxley Act (2002). United States Congress. URL: `https://www.sec.gov/about/laws/soa2002.pdf`.

Scaffidi, C., A. J. Ko, et al. (2006). "Dimensions Characterizing Programming Feature Usage by Information Workers." *Proceedings of the Symposium on Visual Languages and Human-Centric Computing*, 59–64.

Scaffidi, C., M. Shaw, et al. (2005). "Estimating the Numbers of End Users and End User Programmers." *Proceedings of the Symposium on Visual Languages and Human-Centric Computing*, 207–214.

Simmhan, Y. L. et al. (2005). "A Survey of Data Provenance in e-Science." *ACM SIGMOD Record* 34 (3), 31–36.

Winston, W. L. (2001). "Executive Education Opportunities." *OR/MS Today* 28 (4).

Yin, R. K. (2014). *Case Study Research: Design and Methods*. 5th Edition. Sage Publications. ISBN: 1483302008.

Zwicky, F. (1969). *Discovery, Invention, Research - Through the Morphological Approach*. Toronto: The Macmillian Company.