Spreadsheet Practices and Challenges in a Large Multinational Conglomerate

Justin Smith, Justin A. Middleton
North Carolina State University
Raleigh, North Carolina, USA
{jssmit11, jamiddl2}@ncsu.edu

Nicholas A. Kraft
ABB Corporate Research
Raleigh, North Carolina, USA
nicholas.a.kraft@us.abb.com

***Abstract*—Spreadsheets are ubiquitous. Thus, it is important to understand the challenges faced by spreadsheet users in practice. To better understand these challenges, we surveyed ABB employees and then interviewed a cross-section of survey respondents. We used a two-phase coding process to classify the challenges they described. Our survey findings demonstrate that practices in our single-company setting are consistent with practices in broader settings. Our interviews revealed both individual and organizational challenges. For instance, individual participants described data pipeline challenges related to importing data from external sources or storing and archiving spreadsheet data. Further, participants' collective responses revealed challenges pertaining to knowledge distribution within the organization. We outline possible interventions to address these challenges. Our results will help guide researchers and tool designers in addressing the practical challenges facing spreadsheet users.**

I. Introduction

Spreadsheets are one of the most commonly used programming tools in practice [1] and, accordingly, are important to study. Much like traditional programming environments, spreadsheets allow programmers to store data and write code (i.e., formulas and macros) to manipulate that data.

The software engineering community has recognized the parallels between traditional programming environments and spreadsheets [2]. As summarized by Ko et al. [3], researchers have been working to broadly improve spreadsheet practices by applying software engineering techniques such as debugging, refactoring, testing, and fault localization. However, despite decades of research, errors persist and organizations remain overconfident in the correctness of their spreadsheets [4].

As we will argue, errors may persist because we do not fully understand the fundamental challenges faced by spreadsheet users. Our goal is to identify those challenges and the practices that perpetuate those challenges. Prior work provides initial insights into these challenges via studies focused on specific practices, such as testing [5], or on specific populations, such as students [6] or individual spreadsheet users [7].

Because spreadsheets do not exist in isolation — they often become organizational assets shared between individuals [8] — we study spreadsheet usage in a practical setting, *within* an organization (ABB). Our motivation for studying a large multinational conglomerate (i.e., a large and diverse organization formed via both organic growth and acquisitions) is to determine whether the practices within a single ecosystem are consistent with those reported across many organizations in previous studies and to understand the challenges that users face within a large and diverse ecosystem. We frame this study around the following three research questions:

RQ1: What practices are common among professional spreadsheet users?

RQ2: Are the reported practices consistent with previous findings on spreadsheet practices?

RQ3: What challenges do professional spreadsheet users face?

To answer these research questions, we conducted a mixed methods study of spreadsheet practices at a large multinational conglomerate. First, we conducted a survey of 180 spreadsheet users across various roles at ABB (Section IV). We then compared the results of this survey with previously reported studies to assess the representativeness of this sample and contextualize our findings. Next, we conducted 22 follow-up interviews to understand the specific challenges spreadsheet users face (Section V). From those interviews, through a qualitative coding process, we identified 13 challenges. Because our survey results are consistent with the results of previous studies, we believe that our interview results — and thus the challenges we identified — are applicable to other organizations.

In summary, this paper presents the following contributions:

- A description of a spreadsheet ecosystem (Section IV-B).
- A categorization of challenges faced by spreadsheet users (Section V-C).

This work serves as an overview of spreadsheet practices and challenges, which will serve researchers and tool designers in their future work. As we will describe, those challenges exist on both the individual and organizational levels. For example, individuals may face challenges importing data from a content management system to a spreadsheet, while an organization may face challenges producing assets that perform this task across individual business units.

II. Related Work

In this section, we first review the settings of the studies to which we compare our survey findings. We then discuss other work related to identifying spreadsheet challenges.

978-1-5386-0443-4/17/$31.00

A. Surveys about Spreadsheets Practices

Spanning back decades, researchers have surveyed spreadsheet users to better understand the challenges of spreadsheets. Throughout the 1990s, several studies examined the lack of control and policies for spreadsheets among MBA students [9], statistics bureaus [10], and other firms [11] [12]. Several subsequent surveys reported similar issues over the following decades while reporting advanced details on control and testing methods, as Pemberton and Robson [13] did when broadly exploring the role that spreadsheets had assumed in business. The quality of spreadsheet testing has particularly drawn attention, as Panko and Ordway [14] updated the literature and legislative history of spreadsheet testing and fraud detection in light of the Sarbanes-Oxley Act in 2002, and Roy et al. [5] conducted interviews over the development testing practices in over 21 countries. However, previous studies tended to cast their attention over broad ranges of respondents to represent varied populations. We instead focus within the boundaries of a single large company.

The surveys most related to our own were conducted by the Spreadsheet Engineering Research Project (SERP) at Dartmouth. To address the entire life-cycle of spreadsheets from design to archiving, the researchers elicited feedback from a range of spreadsheet users. These populations include MBA alumni from two business schools, spreadsheet software firms, and private corporations, totaling 1,597 responses. The results comprise a number of studies probing individual populations [15] and differences in risk awareness [16] and proficiency [8]. Whereas the SERP survey targeted participants from a range of populations, our survey targets a single ecosystem. We want to probe the practices at a single firm, ABB, to discover in depth whether its practices are consistent with those cataloged in prior research.

B. Identifying Spreadsheet Challenges

Ko et al. [3] summarize the area of end-user software engineering, in which spreadsheets research is situated. They highlight several software engineering challenges, such as reuse and testing, that afflict end-users even though these users often lack professional training in these tasks. In other words, they affirm that spreadsheet users are not exempt from the challenges of software design and maintenance. We extend these comparisons in our study, similarly noting many of their challenges while also reporting new ones (Section V-C), such as challenges in data management and manual processes. Ruthruff and Burnett [17] describe six challenges researchers face in adapting fault-localization techniques to end-user debugging environments. Like Ruthruff and Burnett, we are concerned with spreadsheet-related challenges. However, the challenges we identify impact spreadsheet users directly, rather than spreadsheet researchers.

Other studies outline barriers in learning specific tasks. Ko et al. [6] identify six barriers to learning end-user programming systems, and their model describes the phases of a project that are vulnerable to knowledge breakdowns and incorrect assumptions. These problems include when a user cannot conceive solutions to a problem or when a user confronts a tool's unexpected behavior. Chambers and Scaffidi [7] study the learning challenges of spreadsheets as gleaned from forum posts and interviews with educators. The challenges that they identify relate to spreadsheet-specific features (e.g., data visualization or invoking/debugging macros). We extend their challenge model by identifying challenges within a single-organization setting via interviews with practitioners.

Grigoreanu et al. [18] investigate debugging strategies contextualized in information foraging and sense-making frameworks. In their deconstruction of observed user tasks, they draw out the moments in debugging wherein users are most vulnerable to problems, such as when they interrupt their thought processes or skip steps. Additionally, Kulesz and Ostberg [19] extend these investigations into practices of spreadsheet auditing by examining the auditing software itself. They bring attention to the following challenges: presenting meaningful information non-intrusively; handling false positives; and of several others that hinder end-user tools from the perspective of the tool's designer. Although these studies typically involve a small group of subjects to suggest general problems and ours is a broad analysis over a single conglomerate, we share a common goal of improving end-user productivity.

III. Study Context

We conducted this study at ABB, a large multinational industrial conglomerate. ABB is organized into five divisions, including shared group management functions and four global business divisions. At the time of the study, the divisions were organized into 29 heterogeneous business units (BUs) focused on particular group functions, industries, and product categories. Some BUs have been started as new branches of ABB while others have been acquired in buyouts/mergers. As a result, there are relatively few globally-mandated processes across BUs for issues such as spreadsheet usage. Altogether, these BUs employ about 135,000 people and operate in over 100 countries.

ABB employs about 5,000 traditional software developers (i.e., developers who write code in third- or fourth-generation programming languages) but around 100,000 white-collar workers in total. Although we cannot definitively state how many ABB employees use spreadsheets, the number is certainly in the tens of thousands (i.e., an order of magnitude larger than the number of traditional software developers). We located over 300,000 spreadsheet files stored on an internal shared drive. Thus, there is vast potential for business impact by offering improved processes and tools to address the challenges faced by spreadsheet users in ABB. Accordingly, we sought to catalog and understand these users' practices and challenges.

IV. Survey

A. Survey Design

We conducted a survey to understand how ABB employees use, create, and maintain spreadsheets (RQ1). The survey

comprises 20 multiple-choice and 5 free-response questions, which span six spreadsheet-centric themes: usage, creation, testing, documentation, sharing, and training. The survey also includes 10 demographic questions. According to the taxonomy presented by Baldassarre and colleagues [20], our survey is an external, close, improved replication of several previous surveys [21]–[23]. We distributed the survey in June 2016 via Yammer [24] (an enterprise social networking platform used by ABB employees worldwide) and email (to ABB employees who registered for one or more internal Microsoft Office trainings during the first half of 2016). The complete survey is available online [25].

B. Survey Results

In this section we describe our survey results. We first describe demographic data for our respondents. We then highlight findings related to each of the six spreadsheet-centric themes listed in Section IV-A, as well as responses to selected free-response questions. Finally, each subsection concludes with a discussion relating our results to previous studies.

1) Respondent Demographics: We received 180 usable responses to our survey, including 160 complete responses and 20 partial responses missing only the answers to the demographic questions. Among the 160 respondents who provided demographic information, 120 are from North America and 32 are from Europe. Notably, we received 17 or more responses from each of the five divisions of ABB and at least 1 response from 26 of the 29 business units across those divisions. Moreover, respondents' jobs are spread across many functional areas, including sales (11), marketing (8), operations/manufacturing (16), engineering (20), research (9), finance (39), distribution (2), and other (51).

Previous Studies: Like the SERP surveys [8], we primarily draw on North America and Europe for our respondents. The SERP surveys had nearly 10 times more respondents — about 1,600 — spread across many different organizations. Furthermore, other studies have drawn their respondents from social media, mailing lists, MOOCs [5], and magazine subscribers [26], [27]. However, our population is all from a single organization. The distribution of job roles is also similar between our study and SERP. For example, positions in finance represent 30.2% of their survey base compared to our 24.4%. Unlike Caulkin and colleagues who focused on executives and senior managers [28], we focus on a broader array of job roles and functions.

2) Spreadsheet Usage: About 79% of respondents reported having "some experience" or "extensive experience" with spreadsheets, while about 6% reported having "very little experience" and about 15% reported being "very experienced." Thus, we characterize a majority of the respondents as being competent or proficient (as opposed to being novices or experts). Over 75% of respondents reported that they use spreadsheets mainly for analyzing data (e.g., financial or operational data) or tracking data (e.g., budgets, sales, or inventories). No other main purpose was indicated by at least half of the respondents. Overall, the average respondent

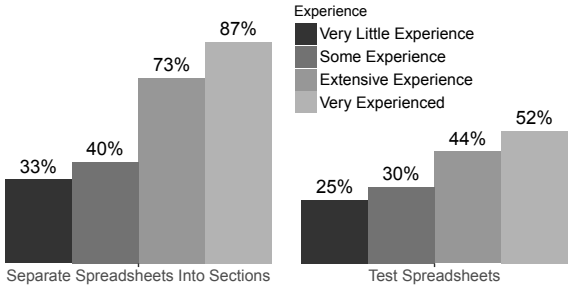

Fig. 1. Survey Results — Percent of users who performed certain tasks more than half of the time, by experience. The chart is based on the responses to questions: "How often do you divide a spreadsheet into separate sections either on a single sheet or by using multiple sheets?" and "When given a spreadsheet to use, how often do you test it?"

reported spending 3–10 hours per week using 2–10 different spreadsheets. Unsurprisingly, 20 of the 34 respondents who reported using spreadsheets for more than 20 hours per week are in finance.

Previous Studies: Our respondents report similar levels of proficiency and expertise as those of Chan and Storey [26]. The primary purposes of data analysis and data tracking are also consistent with the prior SERP survey [8]. Regarding time spent using spreadsheets, our results are similar to those of Lawson et al. [8] and Pemberton and Robson [13].

3) Spreadsheet Creation: Reuse is a fundamental software engineering principle, but 58% of respondents reported creating new spreadsheets from scratch about half the time or more, including 30% of respondents who do so most of the time. Similarly, when creating a new spreadsheet, 55% of respondents enter data and formulas directly, while only 30% borrow a design from an existing spreadsheet. We found no correlation between experience and the typical first step in creating a new spreadsheet, and some "very experienced" respondents indicated that they typically enter data and formulas directly.

Modularization is another key software engineering principle. Spreadsheet users can modularize their spreadsheets by dividing them into separate sections, either across multiple sheets or within a single sheet (e.g., by leaving blank rows or columns between tables). The left side of Figure 1 illustrates that more experienced spreadsheet users are more likely to modularize their spreadsheets than are less experienced users.

Most respondents (about 84%) reported that the spreadsheets they create are normally used by at least one other person, but only 15 respondents (about 9%) indicated that their spreadsheets often become permanent assets in their BUs. Interestingly, 80% of respondents who use macros four or more times per week also share their spreadsheets with many others.

Previous Studies: Previous studies also report that spreadsheet users employ limited reuse strategies. For example, SERP surveys [8] report that 36.3% of their respondents always begin spreadsheets from scratch — 30% of our respondents do so most of the time. The SERP surveys [8] also indicate that modularization comes with experience and that

most spreadsheets are shared.

4) Spreadsheet Testing: Spreadsheets are error-prone [29]. Consequently, understanding (and ultimately improving) spreadsheet testing practices is key, particularly given that spreadsheets users generally lack the knowledge and training of traditional software developers [5]. Our survey results indicate that available tools for spreadsheet testing are not incorporated into spreadsheet testing practices at ABB. Among our respondents, 87% reported examining cells individually or using a calculator to check selected cells, and 79% reported using common sense. Meanwhile, only 40% reported testing performance for plausibility or testing extreme cases, and only 18% reported using Excel-provided testing tools like the data validation and formula auditing tools. Indeed, follow-up interviews revealed that some respondents who reported using Excel-provided tools actually meant using in-built Excel functionality (e.g., summing rows and columns) and were unaware that Excel provided dedicated tools for testing spreadsheets.

The right side of Figure 1 illustrates that more experienced spreadsheet users are more likely to test their spreadsheets than are less experienced users. Nevertheless, only about half of "very experienced" respondents reported testing their spreadsheets more than half the time. Moreover, as discussed in the previous paragraph, few respondents reported using anything but the most basic of testing practices.

Previous Studies: Previous studies also noted the widespread, ad-hoc attempts to manually test spreadsheets [5], and the relative scarcity of systematic testing practices [30]. Roy et al. [5] indicate that testing in practice is common: 99% of respondents do some form of it. Among our population, 81% report testing. The difference is unsurprising given the trend shown in Figure 1 and that their population included more experienced spreadsheet users. Roy et al. report that the most popular technique was ad-hoc testing, performed by 88% of users, which is also consistent with our results.

5) Spreadsheet Documentation: Documentation is key for understanding the design and intent of a software artifact. Over 62% of respondents indicated that they rarely or never document their spreadsheets, and only 23% of respondents document their spreadsheets more than half the time. Nevertheless, 85% of respondents reported documenting spreadsheets at least some of the time. Respondents who share their spreadsheets with others were not more likely to also document their spreadsheets compared with those who rarely share their spreadsheets.

Previous Studies: Our findings are consistent with Roy et al. [5], who found that 80% of respondents practice some form of documentation compared to our 85%. Upon closer examination, both the SERP surveys [15] and our findings suggest consistent documentation practices are rare.

6) Sharing Spreadsheets: Unlike in Section IV-B3, where 'sharing' refers to distributing a spreadsheet for the purpose of reuse, in this section 'sharing' refers to reporting spreadsheet results (e.g., to a manager or customer). Despite the prevalence of modern business intelligence tools such as Qlik and Tableau, spreadsheets are often used as reporting tools. Thus, business users share spreadsheets with colleagues, managers, and customers. Among our respondents, 75% share such reports by providing parts of or the entire spreadsheet. A mere 11% of respondents who share spreadsheets provide a summary, and only 14% of respondents rarely share any part of a spreadsheet.

Different versions of spreadsheets are necessary to accommodate changing requirements and to archive historical calculations. Version control enables change tracking and supports traceability. Despite the popularity of version control systems like Git and the availability of spreadsheet version control systems [31], spreadsheet users do not formally use version control. Indeed, while 74% of respondents typically have two or more versions of a spreadsheet, 64% of respondents track different spreadsheet versions by embedding dates or version numbers in the filenames. Only 4% of respondents save on OneDrive [32], despite OneDrive including rudimentary, but functional, version tracking and being available to ABB employees worldwide.

Previous Studies: Our findings about the most common sharing practices (i.e., sharing an entire workbook) are consistent with the results of the SERP survey [15]. Hermans and colleagues also report that the majority (85%) or spreadsheet users transfer entire spreadsheets to colleagues [33]. In addition, our results on version control are comparable to rates reported by other studies. For example, Roy et al. [5] estimate the use of manual version control among spreadsheet users at 70% compared with our 74%.

7) Free-Responses: We asked participants to list any sources of information that they use for help while creating or editing a spreadsheet. More than half of the respondents (92) listed unofficial online resources such as Google (65), various forums (30), YouTube (16), and Excel-focused blogs (5). About one-third of respondents (58) listed the official Excel help, which is a surprisingly small proportion of respondents, particularly given that we mentioned "Excel help" as an example in the question. Only 33 respondents listed "colleagues" as a source of information, and even fewer respondents listed print media (7), online training (4), previous spreadsheets (3), or official ABB training resources (4).

> Our survey of a single large multinational industrial conglomerate replicates the findings of previous studies that span many organizations. We found that the use of software engineering practices, such as reuse, testing, documentation, and version control, is limited among spreadsheet users. However, our findings also suggest that the use of practices such as modularization and testing becomes more common as users gain experience. Overall, our findings indicate that our population is consistent with the aggregate population considered by previous studies.

V. Interviews

A. Interview Protocol

We recruited participants (Table I) from the pool of survey respondents. All interviews were conducted within two months of the surveys. The pre-interview survey gave us insight into each participant's individual spreadsheet practices and allowed us to tailor the interview to that participant's survey responses.

TABLE I
INTERVIEW PARTICIPANTS

ID	Spreadsheet Experience	Programming Experience	Job Area	Location
P002	●●●●	●●●●○	Research	Switzerland
P018	●●●●	●●●○○	Finance	USA
P025	●●○○	●●●○○	Sales	USA
P026	●●○○	●●●○○	Management	Canada
P029	●●○○	●●●○○	Finance	USA
P030	●●●●	●●●○○	Finance	USA
P035	●●●○	●●●○○	Management	USA
P036	●●○○	●●●○○	Service	Germany
P040	●●○○	●○○○○	Engineering	New Zealand
P048	●●●○	●●●●○	IT	Denmark
P061	●●○○	●●●○○	Distribution	Denmark
P084	●●●●	○○○○○	Finance	USA
P092	●●○○	●○○○○	QA	USA
P095	●●●○	●●●○○	IS	USA
P149	●●○○	○○○○○	Finance	Mexico
P154	●●●○	●○○○○	Finance	USA
P161	●●●●	●●●○○	Marketing	USA
P166	●●○○	●●●○○	Sales	USA
P171	●●●○	●●○○○	Marketing	USA
P181	●●●●	●●●●○	Sales	USA
P193	●●○○	●●○○○	Management	Canada
P202	●●●○	●●●●●	Finance	USA

We invited all 57 survey respondents who provided a valid email address to participate in the interviews. We scheduled interviews with the 25 who responded to our recruitment email. Three participants were unable to attend the interview or reschedule. As a result, we conducted a total of 22 interviews.

The first and third authors jointly conducted each interview over Skype audio calls. Before each interview, participants were asked if their audio data could be collected and stored; all agreed to the recording.

The interviewers followed a semi-structured interview script which included questions about participants' backgrounds, practices, and challenges (full script available online [34]). During the interview, we also discussed participants' survey responses, including their free-responses and any spreadsheets they might have shared. The mean duration for the interviews was 26 minutes and 40 seconds.

B. Interview Analysis

First, we transcribed the audio recordings using oTranscribe [35]. We performed two rounds of coding on each transcript — *initial coding* and *focused coding*. In the remainder of this section we will detail each of these processes.

Initial coding To derive our initial set of codes, we jointly coded two participants' transcripts. The remaining transcripts were divided among the researchers and coded in parallel. Throughout this process, we iteratively updated the set of codes as new codes arose.

To ensure consistency across researchers, all new codes were reviewed by all researchers. Furthermore, to verify codes were applied accurately, each transcript was reviewed by a second researcher. Overall, the two coders agreed on 99% of codes and, in all 11 cases of disagreement, arrived at agreement after brief discussions. This process resulted in the application of 20 codes to 1042 statements.

TABLE II
SUMMARY OF CHALLENGES

Category	Description
Data Sources	Entering/importing data from external sources
Reuse	Solving the same problem repeatedly
Data Management	Storing and archiving spreadsheet data
Out of Practice	Relearning infrequently-used features/techniques
Sharing	Crafting spreadsheets others can understand/use
Complex Applications	Managing a complex domain problem
Infrastructure Limitations	Problems with IT/IS infrastructure or Excel
People	Problems such as inexperience or turnover
Help Resources	Help resources are missing/inaccessible/unhelpful
Features	Difficulty with macros, pivot tables, etc.
Time/Budget	Lack of sufficient resources
Correctness	Difficulty checking correctness of all cells
Manual Processes	Completing tasks manually

Focused coding As we are particularly interested in participants' spreadsheet-related challenges (RQ3), we performed a second round of *focused coding* on the statements that were assigned the "Challenge" code by the *initial coding* phase. In total, the *initial coding* phase produced 195 such statements.

Similar to before, we derived our initial set of challenge codes by jointly coding transcripts. After coding the challenge statements from three participants, we assigned the remaining statements to be coded independently. Any new codes that arose during independent coding were reviewed jointly. Again, to ensure the accurate application of each code, all independently assigned codes were reviewed by a second researcher. Three statements (2%) were disputed as a result of this review. As a result of this *focused coding* process, we identified 13 challenges, each of which we will discuss in the following section.

C. Interview Results

In this section, we discuss each of the challenges identified during our interviews. Each subsection corresponds with one of the challenges and includes a short description of that challenge. To contextualize these challenges, each section also includes selected quotes.

Next to the title of each subsection, we report the number of participants who identified with that challenge in parenthesis (X). Although we report how often each of these challenges was mentioned, we make no quantitative generalizations about the frequency or impact of these challenges. We suspect some of these challenges might resonate more or less within an organization. Instead, we focus on describing each challenge and the avenues through which it might make an impact.

1) Data Sources (12): Spreadsheets are not the only tool participants used to store and analyze data. Spreadsheets exist within a complex landscape of other data management solutions. This challenge relates to issues participants expressed moving data into spreadsheets and formatting that data for subsequent processing. Some participants faced challenges even getting data imported into Excel. P025: "I can't ever seem

to get the data that I'm looking for. It's either not in table form or the Excel import function doesn't read it in table form."

P095 impressed upon us how reformatting input from various data sources can be time consuming: "Whenever I need to do analysis in Excel, I spend 50 percent of my time, half of my time, just cleaning up the data so I can actually use the data in Excel in a meaningful way. That's probably my biggest challenge."

2) Reuse (6): Even across diverse projects, software engineers reuse components — from common data structures to library functions. On the other hand, participants did not describe a similar reuse culture surrounding spreadsheets. Instead, participants, not knowing where to find reusable components, were forced to "start from scratch" when designing spreadsheets. P002: "Everyone has built something on their own fitting exactly their situation, their environment, their needs. Of course, nothing fit what I needed to do. So, I started again, from scratch! That was what I'm talking about reinventing the wheel."

3) Data Management (16): Spreadsheets can accommodate a variety of data, but managing that data poses challenges. P026: "There are so many methods and steps you can do to manipulate the data, but getting it to do exactly what you want can be tricky." For some participants, like P018, these challenges were most apparent when working with specific data types (e.g., "timestamps and dates").

As complexity increases, managing data becomes an increasingly costly challenge because functions become difficult to maintain. P095 notes: "I've seen some spreadsheets that are absurdly complex with multiple, multiple lookup values and other crazy functions. I've seen people sitting down for hours updating their Excel sheet."

P035 elaborates, colorfully describing the resulting spreadsheets: "You look at the formulas and it's just a blob of parenthesis. Your scripts start looking like someone took a chicken and dumped her feet in ink and stamped it on the page. It made sense at one time. Now, you don't know what."

4) Out of practice (5): Some participants use spreadsheets and certain spreadsheet features intermittently. For example, participants described using particular reporting spreadsheets monthly or quarterly. The irregularity of their work lead to the challenge of falling out of practice. P166 succinctly described this challenge, "I learn and relearn a lot." P035 elaborated on the sentiment: "You use it and figure it out today. Two months later, you run into the same thing and just can't remember. It's like 'How the heck did I deal with this same issue?' "

Participants' sporadic use of spreadsheets caused challenges in several situations, ranging from formatting data to rounding results. For P154, this challenge recurringly complicated one of his/her common tasks: "I deal with millions of dollars in terms of numbers often. So, I wanted to round to the nearest million, but couldn't remember for the life of me how to do it."

5) Sharing (12): Individuals and business units use spreadsheets to track data and perform calculations. That data and the results of those calculations often need to be shared with others in the organization. Challenges arise when spreadsheets are not shared in a digestible format. Unfortunately, choosing the correct format depends on the audience and opaque organizational norms.

On one hand, some participants, like P018, preferred sharing spreadsheets in a malleable format (i.e., the sheet itself rather than screenshots or summary reports): "We have some entities that don't send us the Excel sheet. They will take a screenshot of it, which isn't helpful at all."

On the other hand, the rule of sharing the spreadsheet itself does not apply in all circumstances. P095 explained that some audiences prefer to receive summary reports rather than the full spreadsheet: "If you are a senior manager, you don't want to open up an Excel spreadsheet. You want to open up a webpage or PowerPoint."

6) Complex applications (8): Similar to the issues that arise in software product line engineering [36], spreadsheet users face challenges due to the inherent complexity of their work domains. Storing and modeling complex data necessitates complex spreadsheets. For some participants, like P040, these complex spreadsheets evolved over time: "[The product] was only supposed to have about 6 variants. We decided to put the variants in a spreadsheet so we could select them. Then it ended up the customers really liked the product, but they wanted different variants. We ended up in our first spreadsheet of over 3000 variants."

7) Infrastructure Limitations (9): Infrastructure limitations (e.g., hard-drive space, computing resources, limited bandwidth) posed challenges to the participants we interviewed. Participants most commonly faced these challenges when working with large data files. P029: "Because of the sheer volume of information and a lot of formulas that reference that information, my Excel was crashing because it just was not able to keep up."

Further, P036 expressed frustration with the limited bandwidth at his remote site and the lack of compatible training resources. His slow connection prevented his team from simultaneously accessing Excel training videos: "When two people look at the videos there are some seconds shown... Then you go to plant your coffee, grow the coffee up, roast the coffee, then cook the coffee. When you come back you can see the next two or three seconds." Elaborating, P036 explained that alternative training resources, like text files, would be helpful.

8) People (12): Individuals bring a variety of backgrounds, feature preferences, and domain knowledge to any project. As it pertains to projects using spreadsheets, some individuals may have years of professional programming experience and prefer creating macro-enabled spreadsheets, whereas others prefer to use a narrower set of features. These differences between individuals give rise to challenges, especially when ownership of a spreadsheet is transferred from one person to the next.

Participants described how they (or other maintainers) possessed unique knowledge of their organization's spreadsheets. P149 puts it simply: "I don't think anyone else knows this." Unfortunately, this poses a risk of knowledge loss when a

spreadsheet changes hands. In fact, P166 explained to us how he was coping with such a knowledge loss: "There's a bunch of stuff going on in [the spreadsheet] that [my former coworker] created... He's gone now. Now I'm fumbling around trying to figure out how to keep that thing going."

9) Help Resources (12): Despite the overabundance of help resources and training materials available, participants described difficulties applying them to their work at ABB. Due to the specificity and proprietary nature of their work, participants faced challenges finding help from external sources. P026: "The data we work with can be specific in the context of ABB... There's almost a caveat that your data doesn't apply to what the teaching material shows you." This challenge seems to affect spreadsheet users and professional developers alike; Ford and colleagues [37] identify a related challenge, "Abstraction Process barriers," among software developers seeking help on Stack Overflow.

Participants also encountered challenges locating reliable help resources within ABB. In particular, participants faced challenges effectively navigating their social networks. P035: "Identifying the subject matter experts has been kind of difficult. I've asked a lot and they go 'Hmmm... I don't use that. Ask this person' " This challenge is further magnified for more experienced spreadsheet users like P084: "Finding help gets harder the more you know!"

10) Features (15): Spreadsheet software offers many complex features. As participants stressed, making effective use of all of those features can challenge an individual. Moreover, coordinating how those features are used across individuals can challenge an organization.

Usage of any given spreadsheet feature depends on an individual's preferences and experience. Participants, like P166, explained how it can be difficult to gain exposure to new features: "If you don't know what the capabilities are of Excel, you don't even know what to type into Excel to figure it out." Even when participants were aware of advanced features, they often hesitated to use them, citing challenges with maintainability. P035 described this trade off between feature elegance and maintainability: "I can add a lot of great features, but I have to be careful to just add things that I actually need, because I also have to maintain this. I don't want to create my own headache!"

11) Time (8): Of course, limited time, money, and other resources is not a unique challenge to spreadsheet users. Unsurprisingly, participants mentioned these challenges in their interviews. P025: "We often have to make these sheets up in a flash. I mean, we don't have a lot of time to prepare most of what we do here."

Due to the global nature of the organization, participants faced additional time challenges. Some participants, like P166, found it difficult to take advantage of company-sponsored live training: "The training is usually 3:00 or 4:00 in the morning. I work late and am not into getting up that early for the training."

12) Correctness (6): Like any other piece of software, spreadsheet programs can contain syntactic and semantic errors. Participants expressed concern over these errors and ensuring their spreadsheets produce correct results. This challenge was magnified by the perception that mistakes are easy to make. P061: "Well, the problem with spreadsheets is you can easily type something wrong into it, or delete something that you are not supposed to delete, or change something you are not supposed to change. That creates some doubts and mistakes."

Furthermore, once those errors have been inserted into a spreadsheet, they can be difficult to find and remove, especially when spreadsheets are not reviewed. P095: "It makes me nervous when nobody else is checking to say, 'No, now that doesn't make any sense.' In that particular case, if I had forgot that step, or did that step wrong, I'd be reporting incorrect data…"

13) Manual Processes (6): Spreadsheets enable users to build large-scale programs with many tasks automated. However, participants described instances when they encountered the limits of automation. At those limits, participants resorted to completing their tasks manually by 'doing things over and over again' and 'manually massaging' data.

Participants, like P061, found such manual work time consuming: "My experience is that you could end up using quite a lot of time in your spreadsheet doing different kinds of things over and over again. Learning how to make it easier for yourself would be a priority."

Others, like P149, described manual work as error prone: "I have to be very careful because sometimes I forget to copy all the lines that I need."

VI. Discussion

In this section, we discuss how challenges from the preceding section relate to each other and then provide insights on how to address the challenges. We organize our discussion of these synthesized challenges thematically into two topics: *data pipeline* issues and *knowledge distribution* issues.

Data Pipeline: Across several categories, participants described challenges pertaining to what we will call the *data pipeline*. Together these challenges relate to inputting data into spreadsheets (Section V-C1) and managing that data once it arrives in a spreadsheet (Section V-C3), including using burdensome manual processes to manipulate that data (Section V-C13).

As participants described to us, a typical data pipeline might pull data from multiple sources such as SAP, internal data warehouses, and other spreadsheets. Once in a spreadsheet, the user may use some combination of macros, formulas, and manual manipulation to merge these data sources and to transform input to the desired format. Finally, after performing calculations, the resulting data may be used to make business decisions, shared with others, fed as input to another data pipeline, or archived in a database or other data store.

Both individuals and organizations feel the impacts of challenges in these interconnected data pipelines. Individuals who require macros to import and manipulate spreadsheet data are susceptible when changing requirements precipitate modifications to those macros. The most persistent individuals

may seek help or tinker with the macros to patch their own pipelines, but as we have described, these activities are prone to their own challenges. Organizations feel these challenges when it comes to abstracting individuals' custom pipelines into reusable tools and components (Section V-C2). Across individuals, organizations must build infrastructures that support spreadsheet users in managing and moving data.

Outside the domain of spreadsheets, data scientists and software developers face similar issues engineering the data pipeline [38]. However, compared with data scientists and developers, spreadsheet users may uniquely struggle with these challenges, due to their relative inexperience writing scripts and macros. According to our survey results, nearly half of respondents report never having used macros.

To alleviate some of these *data pipeline* issues, we argue that tools should enable spreadsheet users to import and format data using their domain knowledge, without needing to write scripts or macros. Such tools could draw on existing approaches for generating programs from specifications. For example, given a schema for a system like SAP, tools could generate macros that apply standard transformations to auto-populate the spreadsheet fields of interest. This approach would also improve traceability within an organization, helping to propagate updates when changes to upstream data sources are made.

Knowledge Distribution: Individuals with knowledge of particular spreadsheet practices may be dispersed throughout an organization. Spreadsheet users face challenges navigating social and organizational structures to attain the knowledge they need. Here we refer to this issue as a *knowledge distribution* issue. This issue synthesizes several spreadsheet-related challenges, including those pertaining to people (Section V-C8), how they seek help from others (Section V-C9), and how they reuse others' components (Section V-C2).

The issue of knowledge distribution in spreadsheet usage parallels a similar issue in traditional software engineering. Begel and Zimmermann [39] identify this issue as a collaborative information need, phrasing it as a question: "How can we share knowledge more effectively to code faster?" Further, Herbsleb and Mockus [40] attribute delays in globally distrusted development to knowledge distribution issues across weak social networks.

Unlike traditional software developers, spreadsheet users lack the processes, like pair programming [41], and widely-used resources, like StackOverflow [42], necessary to overcome knowledge distribution issues. To illustrate how *knowledge distribution* issues impacted participants, consider the following example: During his interview, P061 described his frustrating experience trying to implement a spreadsheet that automatically calculates an employees' working days — basically weekdays minus holidays and days off. He found this task was particularly challenging for several reasons. For one, different parts of his business unit observed different holidays, requiring him to cross-reference an internal calendar. Despite asking colleagues for help and scouring the web, P061 could not find a solution. Later, we interviewed P029, who shared with us a spreadsheet containing a clever (and well-documented) solution to the exact working days problem P061 had described.

By framing P061's challenges as a knowledge distribution issue, we observe readily-available solutions within his social network. To foster the dissemination and discovery of best practices and help people like P061 find solutions, we propose creating an online community for spreadsheet users within ABB. We will curate this community to highlight annotated examples of best-practices and facilitate pairing struggling practitioners with experts. Because the community will exist within the organization, the material will be specific to ABB employees' needs.

VII. Limitations

Our work faces the threat that we have only identified a subset of challenges facing spreadsheet users and that interviewing more participants would reveal more challenges. However, we reached saturation [43] relatively early — no new challenge codes were introduced after the fifth participant.

Our interviews and qualitative coding approach are also subject to internal threats to validity. To mitigate these threats, we thoroughly reviewed each statement we analyzed; each statement was twice examined by two researchers (Section V-B). Our high rates of agreement (99% & 98%) across coders imply a common understanding of the category definitions.

The survey and interviews were conducted in the context of ABB, which afforded us a richer understanding of the challenges within an organization, specifically the challenges that arose as a result of interactions between individuals. However, this choice threatens our ability to generalize our findings to other spreadsheet users. To mitigate this limitation we sampled participants from geographically and functionally diverse business units. Furthermore, our study included several participants who were new to ABB. Some were either recently hired or joined as a result of businesses acquisitions. Additionally, we compared the practices at ABB to those previously reported (RQ2). Largely, these practices parallel those from prior studies. Altogether, this gives us reason to believe the challenges we identified would apply in other contexts.

VIII. Conclusion

In this work, we report on a mixed methods study of spreadsheet users' practices and challenges. We surveyed 180 spreadsheet users and conduced follow-up interviews with 22 participants from within the same organization. Our results reveal challenges facing individual practitioners as well as cross-cutting organizational challenges, such as sharing and reuse. Through these challenges, we have categorized areas of improvement for spreadsheet tool designers and researchers.

Acknowledgment

We thank Nadeen Saleh, Anthony Benavente, and Michael P. Edwards for their contributions. We also thank the survey and interview participants for their time. This material is based upon work supported by the National Science Foundation under Grant No. 1559593.

REFERENCES

[1] C. Scaffidi, M. Shaw, and B. Myers, "Estimating the numbers of end users and end user programmers," in *Visual Languages and Human-Centric Computing*, 2005, pp. 207–214.

[2] F. Hermans, B. Jansen, S. Roy, E. Aivaloglou, A. Swidan, and D. Hoepelman, "Spreadsheets are code: An overview of software engineering approaches applied to spreadsheets," in *International Conference on Software Analysis, Evolution, and Reengineering*, 2016, pp. 56–65.

[3] 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, "The state of the art in end-user software engineering," *ACM Computing Surveys*, vol. 43, no. 3, pp. 21:2–21:44, 2011.

[4] R. Panko, "What we don't know about spreadsheet errors today: The facts, why we don't believe them, and what we need to do," *CoRR*, 2016. [Online]. Available: http://arxiv.org/abs/1602.02601

[5] S. Roy, F. Hermans, and A. van Deursen, "Spreadsheet testing in practice," in *International Conference on Software Analysis, Evolution, and Reengineering*, 2017.

[6] A. J. Ko, B. A. Myers, and H. H. Aung, "Six learning barriers in end-user programming systems," in *Visual Languages and Human-Centric Computing*. Washington, DC, USA: IEEE Computer Society, 2004, pp. 199–206.

[7] C. Chambers and C. Scaffidi, "Struggling to excel: A field study of challenges faced by spreadsheet users," in *Visual Languages and Human-Centric Computing*, 2010, pp. 187–194.

[8] B. Lawson, K. Baker, S. Powell, and L. Foster-Johnson, "A comparison of spreadsheet users with different levels of experience," *Omega: The International Journal of Management Science*, vol. 37, no. 3, pp. 579–590, 2009.

[9] R. Schultheis and M. Sumner, "The relationship of application risks to application controls: A study of microcomputer-based spreadsheet applications," *Journal of Organizational and End User Computing*, vol. 6, no. 2, pp. 11–18, 1994.

[10] M. J. J. Hall, "A risk and control-oriented study of the practices of spreadsheet application developers," in *International Conference on System Sciences*, 1996, pp. 364–373.

[11] G. Gable, C. Yap, and M. Eng, "Spreadsheet investment, criticality, and control," in *International Conference on System Sciences*, vol. 3. IEEE, 1991, pp. 153–162.

[12] P. B. Cragg and M. King, "Spreadsheet modelling abuse: An opportunity for OR?" *Journal of the Operational Research Society*, vol. 44, no. 8, pp. 743–752, 1993.

[13] J. Pemberton and A. Robson, "Spreadsheets in business," *Industrial Management & Data Systems*, vol. 100, no. 8, pp. 379–388, 2000.

[14] R. R. Panko and N. Ordway, "Sarbanes-oxley: What about all the spreadsheets?" *arXiv preprint arXiv:0804.0797*, 2008.

[15] K. Baker, L. Foster-Johnson, B. Lawson, and S. Powell, "A survey of MBA spreadsheet users." [Online]. Available: http://faculty.tuck.dartmouth.edu/serp/

[16] ——, "Spreadsheet risk, awareness, and control." [Online]. Available: http://faculty.tuck.dartmouth.edu/serp/

[17] J. Ruthruff and M. Burnett, "Six challenges in supporting end-user debugging," in *Proc. 1st Wksp. on End-User Software Engineering*, 2005, pp. 1–6.

[18] V. Grigoreanu, M. Burnett, S. Wiedenbeck, J. Cao, K. Rector, and I. Kwan, "End-user debugging strategies: A sensemaking perspective," *ACM Transactions on Computer-Human Interaction*, vol. 19, no. 1, p. 5, 2012.

[19] D. Kulesz and J.-P. Ostberg, "Practical challenges with spreadsheet auditing tools," *arXiv preprint arXiv:1401.7583*, 2014.

[20] M. T. Baldassarre, J. Carver, O. Dieste, and N. Juristo, "Replication types: Towards a shared taxonomy," in *International Conference on Evaluation and Assessment in Software Engineering*. New York, NY, USA: ACM, 2014, pp. 18:1–18:4.

[21] Spreadsheet Engineering Research Project, "Survey on spreadsheet usage." [Online]. Available: http://bit.ly/SERP_Survey

[22] T. Grossman and O. Ozluk, "Research strategy and scoping survey on spreadsheet practices," in *EuSpRIG 4th Annual Conference Proceeding*, Jul. 2003.

[23] L. Beckwith, D. Inman, K. Rector, and M. Burnett, "On to the real world: Gender and self-efficacy in excel," in *Visual Languages and Human-Centric Computing*, 2007.

[24] Yammer. [Online]. Available: https://www.yammer.com

[25] VL/HCC'17 survey. [Online]. Available: https://figshare.com/articles/Usage_Survey/5113354

[26] Y. E. Chan and V. C. Storey, "The use of spreadsheets in organizations: Determinants and consequences," *Information & Management*, vol. 31, no. 3, pp. 119–134, 1996.

[27] C. Scaffidi, A. Ko, B. Myers, and M. Shaw, "Dimensions characterizing programming feature usage by information workers," in *Visual Languages and Human-Centric Computing*, 2006, pp. 59–64.

[28] J. P. Caulkins, E. L. Morrison, and T. Weidemann, "Spreadsheet errors and decision making: Evidence from field interviews," *Journal of Organizational and End User Computing*, vol. 19, no. 3, p. 1, 2007.

[29] R. Panko, "What we know about spreadsheet errors," *Journal of End User Computing*, vol. 10, pp. 15–21, 1998.

[30] F. Hermans, "Improving spreadsheet test practices," in *Conference of the Center for Advanced Studies on Collaborative Research*, 2013, pp. 56–69.

[31] R. Moreira, *SheetGit: A Tool for Collaborative Spreadsheet Development*. Cham: Springer International Publishing, 2016, pp. 415–420.

[32] OneDrive. [Online]. Available: https://onedrive.live.com

[33] F. Hermans, M. Pinzger, and A. van Deursen, "Supporting professional spreadsheet users by generating leveled dataflow diagrams," in *International Conference on Software Engineering*. New York, NY, USA: ACM, 2011, pp. 451–460.

[34] VL/HCC'17 interview script. [Online]. Available: https://figshare.com/articles/ABB_Spreadsheet_Interview_Protocol/5113360

[35] oTranscribe. [Online]. Available: http://otranscribe.com

[36] A. Metzger and K. Pohl, "Software product line engineering and variability management: Achievements and challenges," in *Proceedings of the on Future of Software Engineering*. New York, NY, USA: ACM, 2014, pp. 70–84.

[37] D. Ford, J. Smith, P. J. Guo, and C. Parnin, "Paradise unplugged: Identifying barriers for female participation on stack overflow," in *International Symposium on Foundations of Software Engineering*. ACM, 2016, pp. 846–857.

[38] T. Barik, R. DeLine, S. Drucker, and D. Fisher, "The bones of the system: a case study of logging and telemetry at microsoft," in *International Conference on Software Engineering*. ACM, 2016, pp. 92–101.

[39] A. Begel and T. Zimmermann, "Analyze this! 145 questions for data scientists in software engineering," in *International Conference on Software Engineering*. New York, NY, USA: ACM, 2014, pp. 12–23.

[40] J. D. Herbsleb and A. Mockus, "An empirical study of speed and communication in globally distributed software development," *IEEE Transactions on Software Engineering*, vol. 29, no. 6, pp. 481–494, 2003.

[41] L. Williams, R. R. Kessler, W. Cunningham, and R. Jeffries, "Strengthening the case for pair programming," *IEEE Software*, vol. 17, no. 4, pp. 19–25, 2000.

[42] B. Vasilescu, A. Serebrenik, P. Devanbu, and V. Filkov, "How social q&a sites are changing knowledge sharing in open source software communities," in *Computer Supported Cooperative Work & Social Computing*. ACM, 2014, pp. 342–354.

[43] B. G. Glaser and A. L. Strauss, *The discovery of grounded theory: Strategies for qualitative research*. Transaction Publishers, 2009.