Delft University of Technology
Software Engineering Research Group
Technical Report Series

Evaluating Automatic Spreadsheet Metadata Extraction on a Large Set of Responses from MOOC Participants

Sohon Roy, Felienne Hermans, Efthimia Aivaloglou, Jos Winter, Arie van Deursen

Report TUD-SERG-2016-002

TUD-SERG-2016-002

Published, produced and distributed by:

Software Engineering Research Group
Department of Software Technology
Faculty of Electrical Engineering, Mathematics and Computer Science
Delft University of Technology
Mekelweg 4
2628 CD Delft
The Netherlands

ISSN 1872-5392

Software Engineering Research Group Technical Reports:
http://www.se.ewi.tudelft.nl/techreports/

For more information about the Software Engineering Research Group:
http://www.se.ewi.tudelft.nl/

Note: Accepted for publication in the Proceedings of 23rd IEEE International Conference on Software Analysis, Evolution, and Reengineering (SANER) 2016.

Evaluating Automatic Spreadsheet Metadata Extraction on a Large Set of Responses from MOOC Participants

Sohon Roy, Felienne Hermans, Efthimia Aivaloglou, Jos Winter, Arie van Deursen
Dept. of Software and Computer Technology
Delft University of Technology
Delft, Netherlands
{S.Roy-1, F.F.J.Hermans, E.Aivaloglou}@tudelft.nl, J.Winter@student.tudelft.nl, Arie.vanDeursen@tudelft.nl

***Abstract*—Spreadsheets are popular end-user computing applications and one reason behind their popularity is that they offer a large degree of freedom to their users regarding the way they can structure their data. However, this flexibility also makes spreadsheets difficult to understand. Textual documentation can address this issue, yet for supporting automatic generation of textual documentation, an important pre-requisite is to extract metadata inside spreadsheets. It is a challenge though, to distinguish between data and metadata due to the lack of universally accepted structural patterns in spreadsheets. Two existing approaches for automatic extraction of spreadsheet metadata were not evaluated on large datasets consisting of user inputs. Hence in this paper, we describe the collection of a large number of user responses regarding identification of spreadsheet metadata from participants of a MOOC. We describe the use of this large dataset to understand how users identify metadata in spreadsheets, and to evaluate two existing approaches of automatic metadata extraction from spreadsheets. The results provide us with directions to follow in order to improve metadata extraction approaches, obtained from insights about user perception of metadata. We also understand what type of spreadsheet patterns the existing approaches perform well and on what type poorly, and thus which problem areas to focus on in order to improve.**

I. Introduction

Spreadsheets are popular and widely used in industry across all domains. Panko [1] estimates that 95% of US firms use spreadsheets for financial reporting. One of the reasons for the popularity of spreadsheets is that they offer a large degree of freedom to their users regarding the way they can structure their data. However, this flexibility can be a double edged sword, which makes it very difficult for spreadsheet users to comprehend spreadsheets. From previous research [2] we know that spreadsheet comprehension poses a difficulty when users transfer spreadsheets to each other, to auditors for error-checking, and to software developers for migration. Especially that final scenario is difficult as the developers responsible for migration usually do not have extensive domain knowledge.

We assert that in such cases of spreadsheet transfer, a deep understanding of the spreadsheet at hand can be helpful. In previous work, this has been addressed by, for example, extracting class diagrams and dataflow diagrams [3], [2], but to comprehend a class diagram or a dataflow diagram, a user still needs knowledge of those formalisms. Hence, we would prefer to support them with a simple way of comprehending spreadsheets: natural language. Ultimately, our goal is to extract documentation from spreadsheets automatically.

As a first step in the automatic extraction of documentation from spreadsheets, we aim to extract metadata: determine what cells in a spreadsheet are metadata (or: labels) and what cells they describe. Contrary to software systems or databases, where metadata is structured, spreadsheets do not have universally accepted structural patterns, increasing the difficulty of distinguishing between data and metadata, or to retrieve the corresponding mappings between them.

In previous work, two approaches have been developed that perform metadata extraction from spreadsheets: The *UCheck* approach developed by Abraham *et al.* [4] with the goal of error-checking in spreadsheets, and the *GyroSAT* approach developed by Hermans *et al.* [2] with the goal of data-flow visualization in spreadsheets. It is difficult however to determine the usefulness of these two approaches for the goal of documentation generation, since both approaches were never evaluated on a large dataset, and the evaluations did not have user inputs.

In this paper, we address those shortcomings by collecting a large number of responses from the participants of a popular Massive Open Online Course (MOOC) conducted by the second author of this paper. As part of an optional exercise included in the MOOC, the participants were asked to identify metadata in spreadsheets. We analyze this data, and compare performance of both the approaches against the participant responses. As such, this paper addresses the following research questions:

- **RQ1**: **How do users perceive and identify metadata in spreadsheets?** Insights about this can be used to improve or train automatic extraction approaches.
- **RQ2**: **How well do two existing automatic approaches perform compared to the users?** An empirical evaluation can be used to assess if the approaches

can be reliably used for the purpose of documentation generation.

RQ3: **In what type of spreadsheets do the approaches perform well, and in what type of spreadsheets do they have difficulties compared to users?** An analysis can be used to improve the approaches.

The results of our analysis show that:

1) Identification of metadata by users is characterized by traits or patterns. For example, groups of commonly used words - like *Name, Description, Name of Country, Name of day* frequently get identified as metadata by users. Also data located in specific positions inside tables of spreadsheets - like column headers and row headers, tend to get identified as metadata.
2) Compared to the users, the two approaches yield recall values of 34% and 45% indicating the need to be improved further in order to be practically reliable.
3) Specific types of spreadsheet structures pose challenges to both approaches, like nested block structures sharing metadata, and data blocks separated by blank rows. These challenges need to be overcome in order to make automatic documentation generation feasible.

The contributions of this paper are:

- A dataset with over 100,000 user-identified pairs of spreadsheet cells and the metadata that describe them.
- Insights from this dataset about how users identify metadata in spreadsheets.
- An empirical evaluation of two existing spreadsheet metadata extraction approaches on this dataset.
- An analysis of situations in which the two approaches perform well and poorly.

II. Background and Motivating Example

In this section, we illustrate the concept of spreadsheet metadata, present a definition, and provide summaries of the UCheck and GyroSAT approaches.

A. Example

As an example, consider the spreadsheet shown in Figure 1. The selected cell E2 is described by the column header *"Interest Due"*, for customer *"John"*. This example illustrates a simple case in which identification of the metadata is relatively easy.

However spreadsheets offer a large degree of freedom regarding the spatial arrangement of data; and this can result in a more complicated example as shown in Figure 2.

The selected cell G14, outlined in red, represents the *"(Projected)"* values, but in addition to this, it is also described by the *hierarchical label "CURRENT YEAR"'*, and has the row header *"Fees"*.

In this case, the cells have multiple cells acting as metadata for it, and metadata is *hierarchical* (defined in the next subsection).

	A	B	C	D	E
1	Customers	Loan Taken on	Principal	Percentage Rate	Interest Due
2	JOHN	05/03/2014	€ 566,666.00	12	€ 152,580.64
3	ANTOINE	10/03/2014	€ 666,666.00	12	€ 178,410.78
4	MIGUEL	15/03/2014	€ 45,323.00	15	€ 15,068.35
5	KURT	20/03/2014	€ 44,332,211.00	12	€ 11,718,278.68

Fig. 1. Simple table structure in spreadsheet

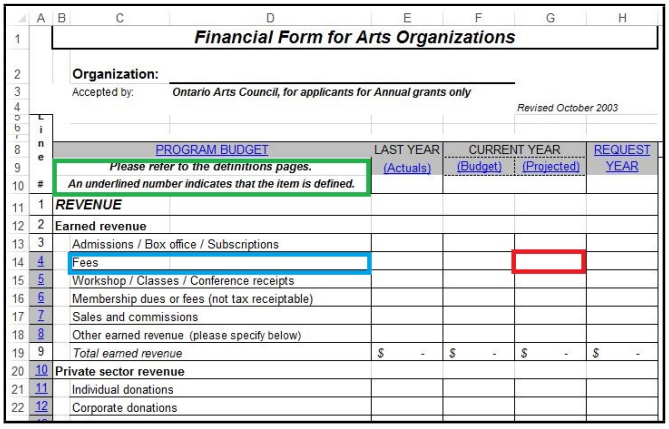


Fig. 2. Complex table structure in spreadsheet: Red cell has 'CURRENT YEAR', 'Projected', 'Fees' as metadata. Blue cell has 'PROGRAM BUDGET', 'REVENUE', and 'Earned revenue' as metadata. Green cell is user instruction

The cell outlined in blue, which itself is metadata, also has *"Earned revenue", "REVENUE" and "PROGRAM BUDGET"* as its metadata. This illustrates the idea that it is possible for metadata to have metadata themselves.

Finally, spreadsheets often contain text that could be considered neither data nor metadata, as seen in Figure 2, where a user instruction in the cell outlined in green is situated in-between the hierarchical metadata cells, and it does not describe any data cell in the spreadsheet.

Figure 2 illustrates a situation where it is difficult to determine what exactly is metadata, due to the complicated spatial arrangement, hierarchical organization, and interspersing of metadata with other types of information.

B. Spreadsheet Metadata

Now that we have illustrated metadata in the two examples above, we present a definition.

In his work related to spreadsheet metadata for the purpose of searching spreadsheets, Chatvichienchai [5] defines metadata as *"Metadata is data about data, more specifically a collection of key information about a particular content, which can be used to facilitate the understanding, use and managementof data."* Thus, spreadsheet metadata is information about data that is stored and manipulated inside spreadsheets. Within the context of this paper, we are only considering information that is available inside spreadsheets themselves; metadata that might exist in the form of documentation outside of the spreadsheet, like a manual in a Word document, is outside scope of this work.

Hierarchical metadata is metadata of metadata, as shown in Figure 2, where the hierarchical order of metadata is *PROGRAM BUDGET - REVENUE - Earned Revenue - Fees*.

In the next subsection we describe the two approaches of metadata extraction evaluated in this paper.

C. Two Approaches for Spreadsheet Metadata Extraction

1) UCheck approach: The UCheck approach [4], [6] was developed by Abraham *et al.* for supporting error checking in spreadsheets based on their unit reasoning system [7]. In order to achieve this, the approach performs spreadsheet metadata extraction. The metadata extraction system developed for this approach, referred to by its authors as the *header inference* system, is an integration framework for four different strategies that are used to classify spreadsheet cells into the categories Header, Core, Footer, and Filler as described in Table I. The system classifies the cells following each of the four strategies. However, since the authors of the system believed that the strategies are not equally accurate in identifying cell types, they allocated confidence levels ranging from 0 (low) to 10 (high) for the classifications based on the respective strategies followed. Therefore after the classifications are completed, if one particular cell gets classified into different categories, the system selects the most suitable category by summing up the respective confidence levels and picking the highest sum. For example if a cell is classified as Header by strategy S1 with confidence level 5, and as Core with strategies S2 and S3 with confidence levels 4 and 2 respectively, then it is classified as a Core cell.

The four strategies used for cell classification are as follows.

- **Content-Based Cell Classification**: Cells are classified based on their contents. For example, cells with aggregation formulas are classified as footer cells, cells with numerical values are classified as core cells, and cells with string values are classified as header cells.
- **Fence Identification** and **Region-Base Cell Classification**: First 'fences' or boundaries of tables are identified and thereafter cells lying on these boundaries are classified with increased levels of confidence due to their position. For example, top-most or left-most cells are classified as headers and lower-most cells as footers.
- **Footer to Core Expansion**: Firstly cells with aggregate formulas are identified and marked as footers. Next the cells that are referred by these footers are marked as core cells, and so is their immediate neighbours if they have the same type of content. In this manner the core region is expanded. Thereafter the left over cells are classified either as header or filler depending on whether they are empty or not.

Once classification of all the cells of a spreadsheet is completed, the header inference system assigns the core cells a row header and a column header. For any particular core cell, the nearest header cell to the left of it and the nearest header cell above it, are assigned as the row and column headers respectively. Apart from this, the header cells themselves are assigned hierarchical second and higher level headers which are inferred based on a set of rules in a recursive fashion. As the end result, core cells (data) in a spreadsheet get associated with two header cells (metadata) at most, and header cells themselves get associated with higher level header cells (metadata of metadata), except for those for which headers could not be found.

2) GyroSAT Approach: Hermans *et al.* [2] developed the GyroSAT approach for the purpose of aiding spreadsheet comprehension through dataflow visualizations. This approach extracts metadata as it is necessary for labelling the diagrams with the name of the entities they represent. In this approach the algorithm for metadata extraction first performs classification of spreadsheet cells into categories Formula, Data, Label, and Empty as described in Table I. However, unlike the UCheck approach, the cell classification process in this case is based on one single strategy. The strategy is inspired by the *Footer to Core Expansion* strategy of the UCheck approach and the algorithm first identifies all cells containing formulas marking them as *Formula*. Next, based on the contents of the formulas, it marks cells that get referenced by the formula as *Data* unless they got already typed as *Formula* in the previous step. Thereafter, it types the remaining cells either as *Empty* if they are empty, or else as *Label*.

Once the classification of cells is completed, the algorithm proceeds to determine *data blocks*. A data block is defined as a rectangle containing a connected group of cells of type *Data, Label, or Formula*. The algorithm identifies a data block by starting with the left-most and top-most non-empty cell in a spreadsheet and successively expanding it to include the horizontal, vertical, and diagonal neighbours until a point is reached when all immediate neighbours of a cell have either been already included in the block or are empty. Thus, in its purpose, this bears some similarity to the *Fence Identification* strategy of the UCheck approach, as both try to determine the boundaries of the tables inside a spreadsheet.

After identification of data blocks, the algorithm assumes that any data cell, say C12, can have two associated labels, one from its column 'C' and one from its row '12', which we refer to as column label and row label respectively. The algorithm also assumes that these labels can be found on the borders of the data block that the cell C12 is contained in. The algorithm starts by inspecting the first cell in column C and if it is of type Label, then it assigns that cell as column label for C12. Otherwise, the algorithm moves down along the column, cell by cell, until it finds a Label type cell. If it encounters cells of type Formula or Data before it finds a label, then it quits the search without returning any cell as column label. It employs a similar strategy starting with the first cell in row 12 in order to identify the row label for C12. As the end result, most data cells in the spreadsheet get associated with two labels (metadata) at most except for those whose labels could not be found.

3) Comparison of the approaches: An important step in both the approaches is to classify cells into different categories based on the nature of their contents. This classification serves as a basis for distinction between data and metadata. The

TABLE I
CLASSIFICATION OF CELLS

Type of Cell	UCheck Approach	GyroSAT Approach	Example in Figure 1
Cell that is used to describe other data inside spreadsheets.	**Header**: The user uses these to label the data.	**Label**: A cell that only contains text, giving information about other cells.	Cells A1 (Customers), B1 (Loan Taken On)
Cells containing formulas that perform calculations.	**Footer**: These are typically placed at the end of rows or columns and contain some sort of aggregation formula.	**Formula**: A cell containing a calculation over other cells.	Cells E2 and E3
Cells that contain data.	**Core**: These are the data cells.	**Data**: A cell filled with data.	Cells C2, C3
Empty cells.	**Filler**: These can be blank cells or cells with some special formatting used to separate tables within the sheet.	**Empty**: An empty cell.	Not present in Figure 1

categories defined in the two approaches are similar to each other, but the respective authors use different nomenclature as shown in Table I [4], [3].

We observe that two different terms have been used to imply spreadsheet metadata. In the UCheck approach the term *Header* is used to indicate spreadsheet metadata. On the other hand, in the GyroSAT approach the term *Label* is used. In this paper, we use the term *Label* to indicate spreadsheet metadata, except when used specifically in conjunction with the UCheck approach.

We also observe that in the UCheck approach, the elaborate cell classification mechanism is the essence. In contrast, the GyroSAT approach uses a single cell classification strategy without confidence levels. Also, the GyroSAT approach concentrates on determination of data blocks, and assigning labels to data cells starting from the boundaries of the data blocks. In the UCheck approach however, assignment of the headers is done by moving outwards from the core cells instead of starting at the boundaries. Nevertheless, the approaches are similar in the fact that they both try to retrieve two labels or headers for data cells, one from the row and one from the column.

D. Existing Empirical Evaluations

Abraham *et al.* tested their Ucheck approach [4] on two sets of spreadsheets; the first set consisted of 10 spreadsheet examples from a book by Filby [8] and the second set consisted of 18 spreadsheets developed by undergraduate Computer Science students.

Hermans *et al.* performed an empirical evaluation of their GyroSAT approach on 50 spreadsheets [3] and compared them to a benchmark manually created by the authors themselves.

III. EXPERIMENTAL SETUP

As demonstrated by the above approaches, there is research interest in extracting metadata from spreadsheets, with the aim of supporting comprehension or to perform validation. However, a clear limitation these papers present is that they have never been validated with a large set of data. In this paper, we address this shortcoming in both papers, by creating a large, user-generated benchmark of labeling data and comparing both approaches against it. To gather labeling data of real-life spreadsheets, we designed an online game in which subjects were asked to select labels for a given cell in a spreadsheet.

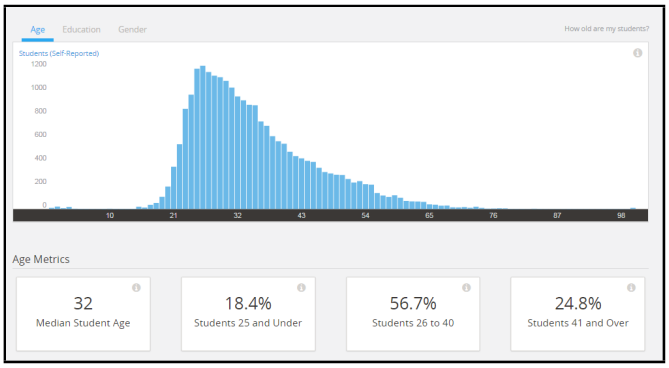


Fig. 3. Overview of age distribution in run 1 of the MOOC

A. Participants

To recruit participants for the labeling game, we included a link to it in the coursework of a popular Excel MOOC: *EX101x: Data Analysis: Take it to the Max()*[1]. The second author of this paper heads the instructor team of this course. The primary goal of the course is to teach participants perform data analysis in general, and work with Excel in particular. The course covers topics like conditional formulas, pivot tables, array formulas and named ranges. It does not however provide any guidance about interpretation or selection of labels for spreadsheet cells, and as such should have no influence on the decisions of the labeling game participants. The course is free and open to everyone, though the target audience are practitioners from various fields who work with spreadsheets often in their daily work. The course also has an optional paid mode offering certificates for identity-verified participants.

To lower the threshold for participating, we did not ask for demographic information of those playing the game, however, we do have the demographics of the entire MOOC: In the two times the MOOC ran, almost 60,000 students participated.

[1]https://www.edx.org/course/data-analysis-take-it-max-delftx-ex101x

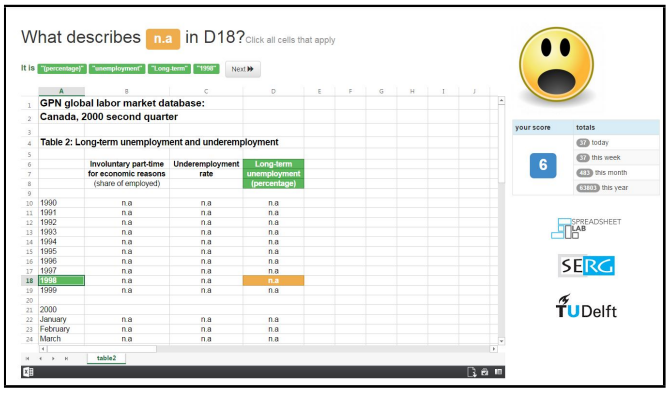


Fig. 4. The interface of the spreadsheet labeling game: Cell D18 is select to be labeled by the user, who has chosen A18, D6, D7 and D8 as labels

The first run of the MOOC started in April 2015, and as shown in Figure 3, the median age of students was 32, and most students (56.7%) fell in the 26 to 40 age group. Almost half of the participants (45.6%) had an advanced degree (MSc or PhD), and a large majority (73.2%) were male. The top countries represented were US (29%), India (11%) and UK (6%).

The rerun in September 2015 was a bit smaller with 23,739 students. The demographics however were similar, with a median age of 30 and 53.5% of students between 26 and 40 years of age; 41.1 % with an advanced degree; 72.3% male students and again US (21%), India (20%) and UK (4%) as top countries.

B. Spreadsheets

As a source of spreadsheets for the game we used spreadsheets from the EUSES corpus [9]. We split up all spreadsheets into separate worksheets, and disregarded worksheets with fewer than 15 non-empty cells, leading to a test set of 1200 spreadsheets. When a user plays the game, they get a random spreadsheet and a random cell to label.

C. The Labelling Game

1) Description: Dubbed 'The Labelling Game' our experiment is presented to users as a game in the browser, as depicted in Figure 4. When playing the game, the user is presented with a spreadsheet in where one cell is highlighted (orange in Figure 4), which we refer to as the *target cell*.

Once the participant has studied the spreadsheet, they can select all the cells that they think describe the target cell, simply by clicking on them. Then, the clicked cells also get highlighted (green in Figure 4) and their contents are recorded as the participant's responses. The participant also has the choice to decline to answer or 'skip' a challenge with the option to record his or her reasons for skipping. Once the participant is satisfied that they have identified all labels for the target cell, they can proceed to the next challenge for a new target cell, and repeat this process for as long as they like.

We attempted to make the labeling game fun by using a smiley displayed in the user's screen as shown in the right of Figure 4. As such there was no 'end' from the perspective of the participants; however to encourage the participants in attempting multiple challenges, the 'happiness' of the smiley was increased with each new target cell they encountered and did respond without skipping. An overview of the number of cells labeled in the past day, week, month, and year was also displayed. The exercise was entirely optional for the course participants and there was no attached benefits promised.

2) Implementation: We implemented the web interface using Javascript and jQuery. For the backend, we use a .NET aspx page accepting json data and writing the results into text files. We use Microsoft's OneDrive and Excel Services JavaScript API[2] to present the spreadsheets to the participants and to collect the participants' cell selections. This API does not, however, support changing the color of cells, which is essential for highlighting the target cell and its user-selected labels. To provide this functionality, we used conditional formatting rules which we included inside hidden worksheets in the spreadsheet workbooks during the pre-processing described in Section III.B.

D. Phases

We ran the Labelling Game in two phases, referred to as the Pilot phase and the Evaluation phase. The Pilot phase was ran in April 2015 during the first run of the *Data Analysis: Take it to the Max()* course, in order to explore the possibility of using such a game for empirical studies. The Evaluation phase was ran during the rerun in September 2015 with some modifications as explained below in order to make it suitable for the evaluation of the UCheck and GyroSAT approaches.

The Pilot phase was intended as a trial in order to gauge the level of involvement from the participants and to assess if such a game could yield sufficient data that can be used for a study. In this phase the target cells were selected randomly at runtime from the set of 1200 spreadsheets used for the game. Thus, the chance of the same target cell being offered to multiple participants was low and we seldom got responses from multiple participants for the same target cell. We realized this was a limitation, as we wanted to establish *correctness* of the identified labels through the number of participants identifying them, or voting mechanism.

We therefore redesigned the experiment slightly during the Evaluation Phase, which was intended for evaluation of the UCheck and GyroSAT approaches. For this phase we manually pre-selected 384 target cells and modified the implementation to randomly pick target cells only from the set of those pre-selected cells. Thus, the probability of the same target cell reappearing to multiple participants was largely increased.

IV. The Dataset

The above described Labelling Game resulted in a large set of data which we describe in this section.

[2]https://msdn.microsoft.com/en-us/library/office/ee589018(v=office.15).aspx

TABLE II
THE DATASET

Description	Pilot phase	Evaluation phase
Total no. of responses	97,526	39,155
Total no. of responses used for study	77,169	30,728
Total no. of target cells in study	26,497	355
Total no. of participants	3040	1183

A. Description

Table II gives an overview of the data, divided in Pilot phase and Evaluation phase. The total number of user-generated responses in the Pilot and Evaluation phases were 97,526 and 39,155 respectively. However, for our analysis we had to discard 21.3% and 21.5% of responses from the two phases respectively due to the following reasons:

1) For our study we analyze the originating spreadsheets from the EUSES corpus, which was not possible for all the cases due to technical limitations, and thus 7.5% and 7.8% of responses in each phase were discarded.
2) The Labelling Game gives the participants the choice to decline from identifying a label by 'skipping' a challenge. Since for our present study we wanted to focus on positively identified labels, we decided to discard such 'skipped' responses amounting to 3.8% and 1.9% from the phases respectively. However, for a future study this subset of responses may make a good candidate for further investigation into understanding what makes it difficult for users to identify labels.
3) To lower the threshold for participating in the game, we did not request for identities of the participants and therefore for practical purposes we decided to assume the IP addresses as unique identifiers for unique participants. In some of the cases IP was not obtained resulting in removal of 0.5% responses in the Pilot phase and 1 response in the Evaluation phase.
4) Lastly, in certain cases we observed single participant identifying an abnormal number of labels for a single target cell. We assumed this type of behavior was due to either misunderstanding the game's objective, or insincerity on part of the participant, and therefore we decided to discard responses tied to all instances where a single user had selected more than 10 labels for one target cell, which amounted to 9.5% and 11.7% in each of the phases respectively.

The total number of responses used for our present study are therefore 77,169 and 30,728 from the Pilot phase and the Evaluation phase respectively.

The total number of randomly picked target cells occurring in Pilot phase was 26,497. The number of target cells randomly selected from the set of 384 pre-selected target cells in Evaluation phase was 355.

The number of participants in Pilot phase and Evaluation phase were 3040 and 1183 respectively.

TABLE III
LABELS FREQUENTLY IDENTIFIED ACROSS SPREADSHEETS

Label	No. of spreadsheets occurring in
1	86
2	72
3	61
5	58
4	54
8	42
Title	42
DESCRIPTION	42
6	40
0	38
2000	37
15	36
10	36
Total	32
year	32
13	31
9	31
Name	30
12	29
7	29
18	26
11	25
14	23
2001	23
16	22

B. Top-3 Ranking and Majority Voting in the Evaluation Phase

A concern regarding the evaluation of the UCheck and GyroSAT approaches was to ascertain the 'correctness' of the labels identified by the participants. One way to address this would be to let multiple participants identify labels for the same target cell and then do a majority analysis on the set of labels for each target cell. In order to achieve this in the Evaluation phase, the target cells were randomly selected by the Labelling Game from a set of 384 pre-selected cells. Consequently, for each target cell T we obtained a set of labels $L = \{l_1, l_2, l_3, ..., l_n\}$ along with their frequencies or votes based on how many participants selected each of them. Therefore it was possible to rank the labels in L based on number of votes.

However, the groups of participants who selected each of the labels in L were not exclusive and had overlaps in-between them. Therefore, we analyzed the Evaluation phase dataset and for each target cell, retrieved the subset of L, say $L' = \{l_1, l_5, l_7, ...\}$ which more than 50% of unique participants had voted for, or in other words, the subset of L that had a majority. We found that for all the 355 target cells in Evaluation phase, in over half of the cases, three labels sufficed to obtain majority. Therefore we decided to evaluate the approaches, as described in Section V, on the top-3 voted labels from the Evaluation phase dataset.

C. User Perception of Labels

We gathered insights about how users identify labels from the dataset we obtained, described in the previous section. These insights could be used to improve existing or develop new metadata extraction approaches. First, we obtained a

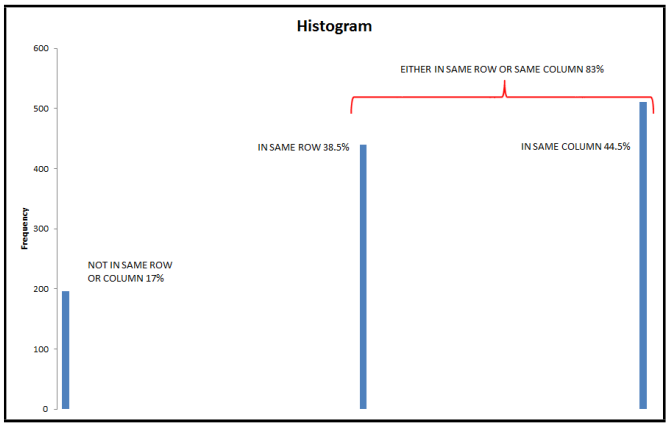


Fig. 5. Location of labels vs. target cells: 38.5% and 44.5%, i.e. 83% in total, of the set of top-3 labels identified by the participants in Evaluation Phase occur either in the same row or in the same column of their corresponding target cell

characterization of the type of words that frequently get identified as labels. Second, we obtained knowledge about spatial localization of identified labels.

1) Frequently used words as labels: In Table III, we have shown 25 most frequently occurring words across multiple spreadsheets that users have identified as labels. This data is provided from the Pilot phase where, as shown in Table II, the number of target cells occurring was much larger, and consequently coming from larger variety of spreadsheets than that in the Evaluation phase. This gives an idea of the words that are commonly identified as labels across 443 EUSES corpus spreadsheets that occurred in the Pilot phase dataset. We observe that:

1. Numbers are often identified as label or metadata. This is a key finding as the automatic approaches usually neglect this aspect, for example in the UCheck approach, the Content-Based classification strategy classifies cells containing numbers as data.

2. Some common words, marked in bold in Table III, like *Title, Description, Total, year, Name, Year 2000, 2001* often are identified as labels. A probable way to use this insight could be the creation of a library of common terms to train the automatic extraction approaches, which is not done as of present in any of the extraction approaches.

2) Role of Row and Column in Labelling: An important aspect to investigate is the spatial location of the labels chosen by the users, as this can provide automatic approaches clues about where to search with more emphasize when attempting to extract labels from spreadsheets. This concept however is already exploited in both the UCheck and GyroSAT approaches. In Section II.C.3 we have re-iterated how the two approaches assume that labels can be found in tuples of two, one coming from the row and one coming from the column, on which the target cell is located. However this assumption in both the approaches was not based upon empirical observation, or validated against user inputs, and thus we wanted to investigate this through analysis on our large dataset from the Labelling Game.

Since both the approaches use this concept as an assumption, we wanted to compare the insight we obtained with the results of the evaluation of the approaches. Hence, we used the set of top-3 voted responses from the Evaluation phase for this analysis, which is the same set used for evaluation of the approaches as described in the next section (Section V) of the paper.

As seen in Figure 5, 38.5% and 44.5%, i.e. 83% in total, of the set of top-3 labels identified by users occur in the same row and same column respectively, as that of their corresponding target cells.

In Section II.C, we have observed how both the UCheck and GyroSAT approaches already have used this insight as an assumption since both try to retrieve labels in pairs, one coming from the row and one coming from the column. This result thus validates that assumption compared to user perception of labels. However, the validity is not lent to the actual results that they yield, which we examine in the next Section V, and which is the combined outcome of all assumptions and followed strategies.

TABLE IV
PRECISION AND RECALL

Phase	Precision		Recall	
	UCheck	GyroSAT	UCheck	GyroSAT
Evaluation phase (top 3)	70%	71%	34%	45%

V. EVALUATION

After having investigated how users perceive labels in spreadsheets, we turn our attention to how the two existing approaches perform compared to users.

A. Accuracy Measures

We compare the two approaches by calculating precision and recall. As explained in Section IV.B, since in over half of the cases in the Evaluation phase, three labels sufficed to obtain the majority vote, we decided to calculate precision and recall against the top-3 of voted answers for each target cell.

In other words, we are calculating whether the two approaches correctly identify the most popular three labels selected by users. The results are shown in Table IV.

1) Precision: Precision measures how many of the labels occur in the top-3 of answers out of the ones the approach retrieved, for each target cell. The average precision is calculated over the whole dataset for all the target cells.

As seen in Table IV, both UCheck and GyroSAT approaches perform fairly well in terms of Precision with average precision of 70% and 71% respectively.

A possible explanation of this can be found in the fact that both the approaches assume rectangular table structures and assume that labels can be found in tuples of two, one coming from the column and one coming from the row. Both the approaches, thus, usually retrieve at most two labels per target cell and therefore evidently are fairly good in terms of precision as compared to users. This is also partly explained by the result (Figure 5) that 83% of the top-3 user selected labels occur in either the same row or the same column of their corresponding target cell.

2) Recall: The value of Recall measures how many of the top-3 labels are selected by the approaches for each target cell. The average recall is calculated over the whole dataset.

As seen in Table IV, the approaches UCheck and GyroSAT do not perform well in terms of Recall with average recall of 34% and 45% respectively.

This result indicates that the approaches are not sufficiently capable of retrieving labels compared to users when performing over a large dataset comprising of real-life spreadsheets. A question that arises at this point is, even though the approaches assume that labels are found either in the same row or column as that of the target cell, which is a valid assumption based on Figure 5, why are they yet unable to provide higher average recall value? The answer to this lies in the fact that the approaches retrieve labels from within the innermost immediately enclosing data-blocks in which the target cell is contained and do not travel across the boundaries of the innermost data-block in search of labels. Yet, our results show, detailed in Section V.B.2, that nested block structures which share a common set of labels across all the blocks are quite common, and are a type of spreadsheet structure that is largely hindering the performance of both UCheck and GyroSAT approaches. Further information on this follows in the next subsection where we explore in detail what type of spreadsheet structures pose difficulties for the approaches.

B. Performance vs. Spreadsheet Structures

In the previous, we evaluated the two approaches in general terms of their accuracy. In this subsection we zoom in more, and explore two different types of spreadsheet structures: those where the approaches perform well, and those where they perform poorly. We are not interested in investigating cases where one approach is performing better than the other as we believe in such cases, one of them has already overcome the other's shortcomings and by combining the two approaches such scenarios can be effectively tackled.

1) Structure Type-I: Both the approaches perform well: Table V shows the top 5 files on which the UCheck approach performed best. We see that on all of them the GyroSAT approach has also faired similarly well, and has performed better in 4 out of the 5 cases. Figure 6 shows the top one in the list *02YEFinSAMPLE.xls* and we see that it has relatively simple structure, similar to the spreadsheet shown in Figure 1, with only one table, no nested data blocks, and no hierarchical headers.

TABLE V
PERFORMANCE BASED ON SPREADSHEETS: GOOD PERFORMANCE

Spreadsheet Filename	UCheck Match Percentage	GyroSAT Match Percentage
02YEFinSAMPLE.xls	60.82%	69.22%
free-excel-tutorial.xls	58.43%	48.34%
Brocade%20GS4%20 Comments.xls	51.91%	72.13%
databasefileonerev.xls	46.63%	55.74%
bb5-list.xls	45.93%	59.28%

TABLE VI
PERFORMANCE BASED ON SPREADSHEETS: POOR PERFORMANCE

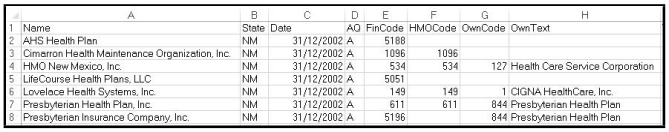
Spreadsheet Filename	UCheck Match Percentage	GyroSAT Match Percentage
2003FinalPopAgeStruct #A857A.xls	0.00%	6.59%
amendment2Section J01a.xls	0.00%	6.78%
Funded%20-%20 February#A835C.xls	9.41%	2.25%
lesson%20planner-soli#A840C.xls	12.33%	0.00%
DCMA.xls	0.00%	13.26%

The approaches perform well with spreadsheets having

- only one table per sheet
- no nested data blocks
- no hierarchical headers

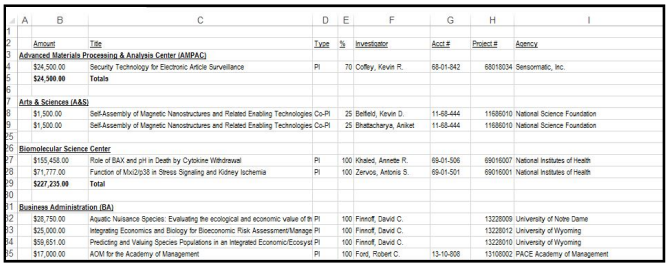
Therefore, in order to create spreadsheets from which automatic extraction of metadata is easier, users can try to adhere to the above mentioned characteristics.

2) Structure Type-II: Both the approaches perform poorly: Table VI shows the 5 spreadsheet files on which both the approaches performed poorly. We see that except for one,



	A	B	C	D	E	F	G	H
1	Name	State	Date	AQ	FinCode	HMOCode	OwnCode	OwnText
2	AHS Health Plan	NM	31/12/2002	A	5188			
3	Cimarron Health Maintenance Organization, Inc.	NM	31/12/2002	A	1096	1096		
4	HMO New Mexico, Inc.	NM	31/12/2002	A	534	534	127	Health Care Service Corporation
5	LifeCourse Health Plans, LLC	NM	31/12/2002	A	5051			
6	Lovelace Health Systems, Inc.	NM	31/12/2002	A	149	149	1	CIGNA HealthCare, Inc.
7	Presbyterian Health Plan, Inc.	NM	31/12/2002	A	611	611	844	Presbyterian Health Plan
8	Presbyterian Insurance Company, Inc.	NM	31/12/2002	A	5196		844	Presbyterian Health Plan

Fig. 6. **02YEFinSAMPLE.xls**: Example of spreadsheet in which both approaches perform well



	A	B	C	D	E	F	G	H	I
1									
2		Amount	Title	Type	%	Investigator	Acct #	Project #	Agency
3	Advanced Materials Processing & Analysis Center (AMPAC)								
4		$24,500.00	Security Technology for Electronic Article Surveillance	PI	70	Coffey, Kevin R.	68-01-842	68018034	Sensormatic, Inc.
5		**$24,500.00**	**Totals**						
6									
7	Arts & Sciences (A&S)								
8		$1,500.00	Self-Assembly of Magnetic Nanostructures and Related Enabling Technologies	Co-PI	25	Belfield, Kevin D.	11-68-444	11686010	National Science Foundation
9		$1,500.00	Self-Assembly of Magnetic Nanostructures and Related Enabling Technologies	Co-PI	25	Bhattacharya, Aniket	11-68-444	11686010	National Science Foundation
10									
11	Biomolecular Science Center								
12		$155,458.00	Role of BAX and pH in Death by Cytokine Withdrawal	PI	100	Khaled, Annette R.	69-01-506	69016007	National Institutes of Health
13		$71,777.00	Function of Mix2/p38 in Stress Signaling and Kidney Ischemia	PI	100	Zervos, Antonis S.	69-01-501	69016001	National Institutes of Health
14		**$227,235.00**	**Total**						
15									
16	Business Administration (BA)								
17		$28,750.00	Aquatic Nuisance Species: Evaluating the ecological and economic value of th	PI	100	Finnoff, David C.		13228009	University of Notre Dame
18		$25,000.00	Integrating Economics and Biology for Bioeconomic Risk Assessment/Manage	PI	100	Finnoff, David C.		13228012	University of Wyoming
19		$59,651.00	Predicting and Valuing Species Populations in an Integrated Economic/Ecosyst	PI	100	Finnoff, David C.		13228010	University of Wyoming
20		$17,000.00	AOM for the Academy of Management	PI	100	Ford, Robert C.	13-10-808	13108002	PACE Academy of Management

Fig. 7. **Funded%20-%20February#A835C.xls**: Example of spreadsheet in which both approaches perform poorly

for the rest either one of the approaches has completely failed to retrieve results. Thus for illustration we choose *Funded%20-%20February#A835C.xls* on which both the approaches have managed to retrieve results but very poorly. As shown in Figure 7, we see that this spreadsheet has similarities with Figure 2 and is characterized by nested vertical data blocks, blank rows separating the blocks, and all the vertical blocks sharing one single set of column headers.

The approaches perform poorly with spreadsheets having
- repeated or nested vertical blocks
- blank rows used to separate the blocks
- vertical blocks all sharing same column headers

The poor performance of the approaches on spreadsheets having above characteristics stems from the fact that both the approaches depend heavily on determination of block structures. UCheck approach uses the fence identification strategy and GyroSAT approach follows the determination of data-blocks based on connected cells. When such blocks are identified, the approaches look for labels in the borders of the blocks, or move outwards from the target cell looking for labels till border is reached. However, as shown in Figure 7, several vertical blocks are repeated, yet they share the same column headers acting as metadata on the top of the spreadsheet. This set of metadata on the top is missed by the extraction approaches when the target cell is located in any of the blocks below that of the first one from top. This is also the reason due to which, as reported in Section IV.C.2, in spite of the assumption that labels occur in the same row or same column being valid compared to user perception, the approaches still fail to perform reliably as the emphasis they put on the immediate surrounding data-blocks, pre-empts their search at the boundaries of such blocks. Therefore for improvement, automatic approaches need to overcome the challenge imposed by such block structures with shared metadata across their boundaries.

VI. Research Questions Revisited

After the analysis of the dataset in Section IV, and the evaluation of the UCheck and GyroSAT approaches in Section V, in this section we revisit our research questions and reflect on the answers.

RQ1: How do users perceive and identify metadata in spreadsheets?

From the results presented in Section IV.C, we can conclude that the way users identify metadata in spreadsheets can be characterized.

Firstly, as shown in Section IV.C.1, we note that numbers are often identified as metadata, a fact which the automatic approaches tend to overlook, as for example the UCheck approach classifies cells containing numbers as data cells based on its Content-Based Classification strategy.

We also observe that certain generic words like *Title, Description, Total, year, Name, Year 2000, 2001* get frequently identified as metadata across multiple spreadsheets. Since it is difficult for automatic approaches to derive any semantic information from contents in a spreadsheet, this finding can prove to be useful if a library is created of frequently used words as metadata. For example, approaches could classify cells containing such words as metadata with higher level of confidence, when using the confidence level technique of the UCheck approach (Section II.C.1). Such libraries can also be created for domain specific terminology to make them more fine grained.

Secondly, as shown in Section IV.C.2, a large majority (83%) of the top-3 labels identified by the users are located in the same row or same column as that of their corresponding target cells. This characteristic is however already utilized as both the UCheck and GyroSAT approaches assume this, and their assumptions are thus validated compared to users.

RQ2: How well do two existing automatic approaches perform compared to the users?

From the results shown in Section V.A, we observe that the UCheck and GyroSAT approaches have average precision of 70% and 71%. We also observe they have average recall of 34% and 45% respectively. From these results we can conclude that although the approaches are fairly precise, they are not practically reliable in terms of capability of retrieval. To be reliably used for documentation generation, a higher recall value is desired. Since both the approaches limit the number of metadata retrieved by 2, a proposition could be to raise this limit to higher values irrespective of the risk of decreasing the precision as precision is already fairly high.

RQ3: In what type of spreadsheets do both the approaches perform well, and in what type of spreadsheets they have difficulties identifying metadata as compared to the users?

As shown in Section V.B.1, we observe that for relatively simple spreadsheet structures with one single table per sheet the approaches perform well. However, as shown in Section V.B.2, for complex spreadsheet structures the approaches fail to perform well. It is observed that nested block structures that share common set of metadata poses problem for the approaches as they only search for metadata around their innermost enclosing data blocks. Therefore it is necessary to develop algorithms that do not limit their search at the boundaries of the innermost enclosing data blocks, but can traverse across boundaries in order to reach the borders of the outermost block or table as well.

VII. Related Work

There are several works related to this research direction. The two approaches under consideration Ucheck [4], [6] and GyroSAT [2] are related, for a more extensive overview, see Section II-C1.

Furthermore there is our own work on the extraction of class diagrams [3] and dataflow diagrams [2]. Cunha *et al.* also worked on extracting information from spreadsheets, with the goal of transforming them into relational databases [10].

Specifically focusing on the spreadsheets made by scientists, de Vos *et al.* [11] have designed a methodology to extract

ontologies in the form of class diagrams from spreadsheets. While their described method is currently manual, they state it could be automated in the future, leading to an interesting new test set for our current work.

Most related is the work by Chatvichienchai, who proposed a spreadsheet layout based metadata extraction approach [5], [12] for the purpose of searching spreadsheets over the web or in document repositories. While the approach shares the goal of extracting metadata, their overarching goal is to return better search results of relevant spreadsheets, causing their metadata to be more high-level than ours.

Chen *et al.* too presented an approach for extracting information from spreadsheets on the web [13] with the goal of integrating spreadsheets with relational database management systems. This approach also performs metadata extraction, but only supports spreadsheets with a simple, *data frame* structure.

While their goals differ, these two final approaches are interesting related works, and in future work we plan to study these too, as they also have not been evaluated against large number of user responses.

Another group of related works concern the usage of MOOC data by researchers. Vihavainen *et al.* used data collected from MOOC participants to successfully introduce techniques for improving participant approval and engagement in a MOOC on programming [14]. Huang *et al.* used data collected from MOOCs to understand behavior of students with increased inclination to post in the MOOC forums [15]. However, this type of research is intended to utilize MOOC data for the education and online education field. In this paper, we have used MOOC data to address the need of large scale user participation in context of empirical software engineering research.

VIII. Discussion

A. Covering Other Approaches of Metadata Extraction

In this paper we evaluated two existing approaches for metadata extraction from spreadsheets. We also created a user generated benchmark to evaluate approaches on. Subsequently, we can evaluate other approaches like [13] against this benchmark as discussed in Section VII.

B. Threats to Validity

1) Threats to External Validity: A threat to external validity of our results concerns the representativeness of the EUSES [9] corpus. However it is a large set, the spreadsheets have been collected from practice, and it has been used in several works of spreadsheet research [16]. In his work Jansen [17] shows how the EUSES corpus is also similar to the more recent ENRON corpus [18], which is a collection of spreadsheets obtained from the e-mail archives of Enron Corporation, disclosed during the trials related to its bankruptcy.

2) Threats to Internal Validity: A threat to internal validity of our results is caused by the manual pre-selection of target cells used for the Labelling Game, during the Evaluation phase. However, completely random pre-selection of target cells results in irrelevant cells being selected, for example blank or empty cells, for which participants tend to 'skip' answering. Thus to obtain more meaningful results, this was a necessary trade-off we opted for.

IX. Concluding Remarks

The objective of this work is to understand how users identify spreadsheet metadata, and how two existing approaches perform compared to the users. The goal is to assess if the approaches can be reliably used as an initial step in automatic generation of documentation from spreadsheets.

In this paper, we have described an experimental setup which consists of an online game included as part of a MOOC. From the large resulting dataset consisting of responses from the MOOC participants, we have learned how users identify spreadsheet metadata, and obtained insights that could be used to improve or develop automatic metadata extraction approaches. In addition, we have also performed evaluation of two existing metadata extraction approaches on the dataset. We observe that the UCheck and GyroSAT approaches of spreadsheet metadata extraction yield average *Precision* of 70% and 71%, and average *Recall* of 34% and 45%, over the whole dataset, indicating the need to be improved further in order to be practically reliable. Specific types of spreadsheet structures pose challenges to both the approaches, like nested block structures sharing same set of metadata, and data blocks separated by blank rows. The results also show that identification of metadata by users is characterized by traits or patterns. For example, groups of commonly used words, and data located in specific positions of tables inside spreadsheets - like column headers and row headers - get frequently identified as metadata by users.

For future work, using all the results and insights obtained from this paper, we aim to develop a spreadsheet metadata extraction approach that can yield better recall compared to the baseline we have obtained for the UCheck and GyroSAT approaches in this study. Addressing the problem of nested block structures, and using a library of frequently used terms as labels for training our extraction approach, are two directions we would like to explore next, towards our ultimate goal of automatic generation of documentation from spreadsheets.

REFERENCES

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

[2] F. Hermans, M. Pinzger, and A. Van Deursen, "Supporting professional spreadsheet users by generating leveled dataflow diagrams," in *Proceedings of the 33rd International Conference on Software Engineering*. ACM, 2011, pp. 451–460.

[3] F. Hermans, M. Pinzger, and A. van Deursen, "Automatically extracting class diagrams from spreadsheets," in *ECOOP 2010–Object-Oriented Programming*. Springer, 2010, pp. 52–75.

[4] R. Abraham and M. Erwig, "Header and unit inference for spreadsheets through spatial analyses," in *Visual Languages and Human Centric Computing, 2004 IEEE Symposium on*. IEEE, 2004, pp. 165–172.

[5] S. Chatvichienchai, "Spreadsheet metadata extraction: A layout-based approach," in *Database and Expert Systems Applications*. Springer, 2012, pp. 147–160.

[6] R. Abraham and M. Erwig, "Ucheck: A spreadsheet type checker for end users," *Journal of Visual Languages & Computing*, vol. 18, no. 1, pp. 71–95, 2007.

[7] M. Erwig and M. Burnett, "Adding apples and oranges," in *Practical Aspects of Declarative Languages*. Springer, 2002, pp. 173–191.

[8] G. Filby, *Spreadsheets in Science and Engineering*. Springer Berlin Heidelberg, 2013. [Online]. Available: https://books.google.nl/books?id=-6HvCAAAQBAJ

[9] M. Fisher and G. Rothermel, "The euses spreadsheet corpus: a shared resource for supporting experimentation with spreadsheet dependability mechanisms," in *ACM SIGSOFT Software Engineering Notes*, vol. 30, no. 4. ACM, 2005, pp. 1–5.

[10] J. Cunha, J. a. Saraiva, and J. Visser, "From spreadsheets to relational databases and back," in *Proceedings of the 2009 ACM SIGPLAN Workshop on Partial Evaluation and Program Manipulation*, ser. PEPM '09. New York, NY, USA: ACM, 2009, pp. 179–188. [Online]. Available: http://doi.acm.org/10.1145/1480945.1480972

[11] M. D. Vos, W. R. V. Hage, J. Ros, and G. Schreiber, "G.: Reconstructing Semantics of Scientific Models : a Case Study," in *In: Proceedings of the OEDW workshop on*, 2012.

[12] S. Chatvichienchai, "Automatic metadata extraction and classification of spreadsheet documents based on layout similarity," in *Advanced Information Management and Service (ICIPM), 2011 7th International Conference on*, Nov 2011, pp. 38–43.

[13] Z. Chen and M. Cafarella, "Automatic web spreadsheet data extraction," in *Proceedings of the 3rd International Workshop on Semantic Search over the Web*. ACM, 2013, p. 1.

[14] A. Vihavainen, M. Luukkainen, and J. Kurhila, "Multi-faceted support for mooc in programming," in *Proceedings of the 13th annual conference on Information technology education*. ACM, 2012, pp. 171–176.

[15] J. Huang, A. Dasgupta, A. Ghosh, J. Manning, and M. Sanders, "Superposter behavior in mooc forums," in *Proceedings of the first ACM conference on Learning@ scale conference*. ACM, 2014, pp. 117–126.

[16] F. Hermans, B. Sedee, M. Pinzger, and A. v. Deursen, "Data clone detection and visualization in spreadsheets," in *Proceedings of the 2013 International Conference on Software Engineering*, ser. ICSE '13. Piscataway, NJ, USA: IEEE Press, 2013, pp. 292–301. [Online]. Available: http://dl.acm.org/citation.cfm?id=2486788.2486827

[17] B. Jansen, "Enron versus EUSES: A Comparison of Two Spreadsheet Corpora," *ArXiv e-prints*, Mar. 2015.

[18] F. Hermans and E. Murphy-Hill, "Enron's spreadsheets and related emails: A dataset and analysis," in *Proceedings of ICSE '15*. IEEE, 2015.

TUD-SERG-2016-002
ISSN 1872-5392