# How plausible is automatic annotation of scientific spreadsheets?

Martine de Vos*
Computer science, Network Institute,
Vrije Universiteit Amsterdam
martine.de.vos@vu.nl

Jan Wielemaker
Computer science, Network Institute,
Vrije Universiteit Amsterdam
j.wielemaker@vu.nl

Bob Wielinga †
Computer science, Network Institute,
Vrije Universiteit Amsterdam

Guus Schreiber
Computer science, Network Institute,
Vrije Universiteit Amsterdam
guus.schreiber@vu.nl

Jan Top†
Food and Biobased Research,
Wageningen University and Research
Centre
jan.top@wur.nl

## ABSTRACT

It is possible to automatically annotate a natural science spreadsheet using lexical matching, given that the tables in these spreadsheets meet a number of requirements regarding the content. Results of a survey show that most of the existing natural science spreadsheets deviate from the ideal situation. We propose to complement lexical matching with both heuristics and knowledge from external vocabularies to overcome these deviations.

## CCS CONCEPTS

• **Computing methodologies** → **Model development and analysis**; • **Applied computing** → **Physical sciences and engineering**;

## KEYWORDS

Spreadsheets, Knowledge engineering, Domain Knowledge, Heuristics, Vocabularies

**ACM Reference Format:**
Martine de Vos, Jan Wielemaker, Bob Wielinga †, Guus Schreiber, and Jan Top. 2018. How plausible is automatic annotation of scientific spreadsheets?. In . ACM, New York, NY, USA, 6 pages.

## 1 INTRODUCTION

In this paper we investigate the feasiblity of automatically annotating existing natural science spreadsheets.

Scientists in the domain of natural science, hereafter referred to as domain scientists, frequently use spreadsheets to analyze and manipulate their research data [13, 18, 25]. The format of spreadsheets gives them a great deal of freedom in how they enter their data. Domain scientists can make their own choices with respect to the entities and processes to be included, and the way in which these are organized in tables. In this way their domain model is implicitly reflected in the content and structure of the spreadsheet tables. As researchers do not anticipate the reuse of their spreadsheet data [23] they tend to be sloppy in the specification of the semantics of their data, and the free format allows them to do so [19]. However, as the domain model is essential to understand the meaning and context of spreadsheet data, it is currently hard to unambiguously interpret these data for people other than the original developers.

The content of the tables can be annotated with concepts from external vocabularies to facilitate interpretation, evaluation and reuse of spreadsheet data. This annotation can be performed automatically by using lexical matching, i.e., by assessing the lexical similarity between spreadsheet terms and labels from selected vocabularies. This method requires, ironically, that the spreadsheet cells contain explicit and complete information on the corresponding research. Furthermore, as natural science spreadsheets often represent observational data, the tables typically contain mostly numbers and little text. The amount of data that can be annotated in natural science spreadsheets is thus limited.

The goal of this paper is to evaluate to what extent automatic annotation of the table content is possible for existing natural science spreadsheets. In Section 3 we explain the requirements for the content of natural science spreadsheets that enable automatic annotation. In Section 4 we present an analysis on the nature and frequency of common design characteristics in existing natural science spreadsheets, and discuss how these deviate from the ideal situation. We propose to repair these deviations by complementing the lexical matching method with heuristics and rules (Section 5) that have been developed in earlier work [6]. Finally, we discuss how these heuristics could be applied to the set of analyzed tables, and to what extent automatic annotation would be possible (Section 6).

## 2 RELATED WORK

Many studies focus on improving the interpretation of spreadsheet data to facilitate reuse and integration. In order to derive a correct interpretation, the studies use different strategies to infer the semantics from spreadsheet data and dissolve ambiguities. We observe two main types of strategies.

One strategy is to encourage domain scientists to standardize their data to facilitate interpretation and reuse. Semantic markup tools like RightField [25] and OntoMaton [13], may be used to develop templates for domain scientists to enter and annotate their

*Corresponding author
†Second affiliation: Computer science, Network Institute, Vrije Universiteit Amsterdam

*K-CAP2017 Workshops and Tutorials Proceedings,*

data simultaneously. MAGE-Tab [18], ISA-Tab [21], and BIOM [14] are tabular formats that use an underlying data model with relevant metadata from scientific experiments. These formats can be used to either directly enter data, or as a template for mapping other spreadsheet files onto one structure.

Another approach is to annotate tabular data with concepts from vocabularies. Some tools [12, 17] use existing generic ontologies, like Yago, DBPedia , while other tools [13, 25], use existing domain ontologies for semantic markup. Some approaches develop their own ontology, either manually [22] or by extracting concepts and relations from the web [24], to annotate tabular data.

All the abovementioned studies acknowledge that a correct interpretation of tabular data is essential for conversion or annotation. In order to derive a correct interpretation, the studies use different approaches to infer the semantics from tabular data and dissolve ambiguities. Some of these approaches rely on manual mapping specifications constructed by users [22] or human analysts with sufficient knowledge of applying semantic web techniques [4, 9, 11, 15, 25]. Others compare their tabular data with large collections of example data, e.g., large vocabularies like Yago or DBPedia, or generic databases extracted from the Web, and rely on probabilistic reasoning methods to find the best suitable annotation or interpretation for table cells and columns [2, 12, 17, 24]. And, many studies use knowledge on the structural properties of a table to derive a correct interpretation of its content. Several studies created a library on commonly used layout patterns in tabular data [8, 10, 20]. Abraham and Erwig [1] developed a framework to automatically classify roles of cells in a table based on the spatial layout of a spreadsheet. Van Assem and colleagues [23] introduced disambiguation strategies for units of measure and quantities ([23]) based on the way these are notated in table cells. And Chen and Cafarella [3] use heuristics and rules on spreadsheet layout and implicit metadata structure to automatically extract relational data from spreadsheets.

## 3 REQUIREMENTS FOR SPREADSHEET TABLES

Spreadsheets from the domain of natural science, e.g., biology, physics and medical science, often represent laboratory or field observations. The tables in these spreadsheets therefore typically consist of numerical data, quantities and units of measure [23], and information on the associated phenomena, i.e., objects, events and substances.

Annotation of these spreadsheets with vocabulary concepts sets requirements for both the vocabularies and the content of the spreadsheet tables. The selected vocabularies should contain labeled concepts, and comprise at least one vocabulary that covers the domain of the considered spreadsheets, and a dedicated vocabulary on quantities and units. In our research we use the OM Ontology for units of Measure and related concepts [19].

Regarding the requirements for the content of spreadsheet tables, the terms representing units of measure should follow the international notation standards [19] (Figure 1). This implies that these terms consist of short strings containing one or more symbols, and optional brackets and slashes [23]. The symbol(s) in the term should lexically match with a unit symbol from the OM vocabulary, , e.g., “ha” representing the unit “hectare” Cells that contain information on quantities should contain a description of a quantity concept, that can be lexically matched with a concept from the OM vocabulary, e.g., “area” or “mass”. Furthermore, quantity cells should have an associated unit of measure, that is located either in the same or in a neighboring cell. Phenomenon cells, i.e. cells containing phenomenon instances, should contain terms that can not be confused with quantities or units, i.e., these terms should preferrably not consist of very short strings, symbols, or abbreviations. The phenomena in tables are annotated with domain concepts, e.g., "corn" and "urea".

**Table 1: Sources of tables inspected to perform the analysis on design characteristics in natural science spreadsheets**

| | Online supplementary data | From institutes |
|---|---|---|
| # research projects | 12 | 8 |
| # spreadsheets | 40 | 44 |
| # tables | 128 | 233 |

## 4 ANALYSIS OF TABLE DESIGN

### Data set

We conduct an analysis on a set of existing natural science spreadsheet tables, in order to gain knowledge on common practice of table design, and to find out in what ways the content of these tables may deviate from the ideal situation as described in the previous section. To this end, we analyse a total of 361 tables in 84 spreadsheets, that are used in 20 existing research projects in the domain of natural science (Table 1). All spreadsheets fall within the scope of our research, i.e., natural science spreadsheets that consist of numerical data, quantities and units of measure, and information on the associated objects and events. About half of the inspected tables is used in our earlier work on interpretation and annotation of spreadsheets [5, 6], or formulas [7]. We collect additional spreadsheets from colleagues at Wageningen University and Research and we use the Google Scholar web search engine to find spreadsheets that are published online as supplementary data alongside journal papers.

### Data analysis

Our analysis consists of a manual inspection of all spreadsheet tables, in which we only consider the content of the tables, and ignore the title or comments. We color code the cells in each block based on the content (see, for example, the legend in Figure 2), and analyze the syntax of the formulas and their composition in the table. Subsequently, we determine for each table to what extent it meets our requirements described in section 3. We explain our results in terms of deviations from these requirements, and discuss some additional observations we made on the structure of the analyzed tables.

### 4.1 Results

*Deviant unit notations.* More than half of the analyzed tables contains unit cells (Table 2), but in almost one third the notation of the unit terms and symbols is not according to the international standards. In the majority of the cases the unit terms are customized

| | A | B | C | D | E |
|---|---|---|---|---|---|
| 1 | | | | | |
| 2 | | | **Corn** | **Cabbage** | **Carrot** |
| 3 | | **Area (ha)** | 30 | 25 | 18 |
| 4 | | **Mass manure applied (kg)** | 150 | 400 | 360 |
| 5 | | **Nitrogen surface density (kg/ha)** | =(C4*C9)/C3 | =(D4*D9)/D3 | =(E4*E9)/E3 |
| 6 | | | | | |
| 7 | | | | | |
| 8 | | | **Coated urea** | **Ammonium nitrate** | **Guano** |
| 9 | | **Nitrogen content (kg/kg)** | 0.4 | 0.25 | 0.15 |
| 10 | | **Phosphorus content (kg/kg)** | 0 | 0 | 0.05 |
| 11 | | | | | |

**Figure 1: Tables in the stylized example spreadsheet**

by the scientists. The resulting unit terms are not incorrect per se, but rather unconventional, and automatic recognition of these terms is hindered. Scientists often combine phenomena with unit symbols, e.g., "MJ/1000 kg milk/yr" and "g CO2e/MJ" (Figures 2, 3).

*Incomplete quantity notations.* The majority of the analyzed tables contains quantity cells (Table 2). In almost half of the analyzed tables, the quantity cells do not contain complete information, and automatic recognition of the quantities is not straightforward. Some tables contain cells with a phenomenon description, and an associated unit of measure located in the neighboring cell (Figure 2). Although no quantity concept is mentioned, these cells implicitly represent quantities. In Table 2 we do not consider these cells as quantity cells.

*Unclear phenomenon notations.* In the majority of the analyzed tables phenomenon cells are present (Table 2), However, part of these tables contains cells that, judging from the position in the table, probably represent phenomena, but do not contain full words. Instead, these cells contain numbers or codes representing, e.g., dates, scientific experiments, identification numbers (Figures 2,3), or abbreviations which are either application specific, e.g., chemical elements or geographical codes, or related to scientific experiments.

*Other observations.* In more than half of the analyzed tables the float cells, containing the values of observations, and the string cells, containing contextual information on these observations, are not located in homogeneous blocks. In most of these tables, the float blocks are interrupted, either by empty cells (Figure 3), or less frequent, by qualitative, i.e., string values. In a small part of the analyzed tables the string and float blocks are not aligned with each other, i.e., these blocks do not have similar dimensions. In these tables it is not clear which observations are associated with which context.

Cells representing semantically related phenomenon instances are typically grouped in the same string block. We observe that these semantic relations are often constructed by the spreadsheet developer. The developer groups instances according to common properties, which may be clear to users or peer scientists, but not easily recognized in a domain vocabulary.

**Table 2: Presence of unit, quantity and phenomenon cells in the analyzed tables**

| Cell type | Fraction of tables (%) | | |
|---|---|---|---|
| | Unit | Quantity | Phenomenon |
| Present, complete info. | 29 | 15 | 46 |
| Present, incomplete info. | 29 | 47 | 29 |
| Not present | 42 | 38 | 25 |

## 5 COMPLEMENTARY HEURISTICS

We observe that most of the analyzed tables do not meet one or more of the requirements listed in Section 3, and we expect that lexical matching will not yield many useful annotations. In this section we therefore propose to complement lexical matching with heuristics and knowledge from vocabularies to overcome the challenges of incomplete information.

### 5.1 Recognizing and annotating blocks

Domain scientists typically group cells that are semantically related [16] and use structure and layout features to distinguish between these groups [3]. We assume that this grouping not only applies to phenomenon cells, but also to cells representing quantities and units of measure.

We have developed heuristics that support us in the recognition of the type of cell, i.e., unit of measure, quantity or phenomenon, and the annotation of its content [6]. These heuristics combine information on the notation of terms in cells, and the composition and positioning of blocks of cells in a table, e.g.:

- *If a cell contains both a string term and a unit of measure, it is a quantity cell*

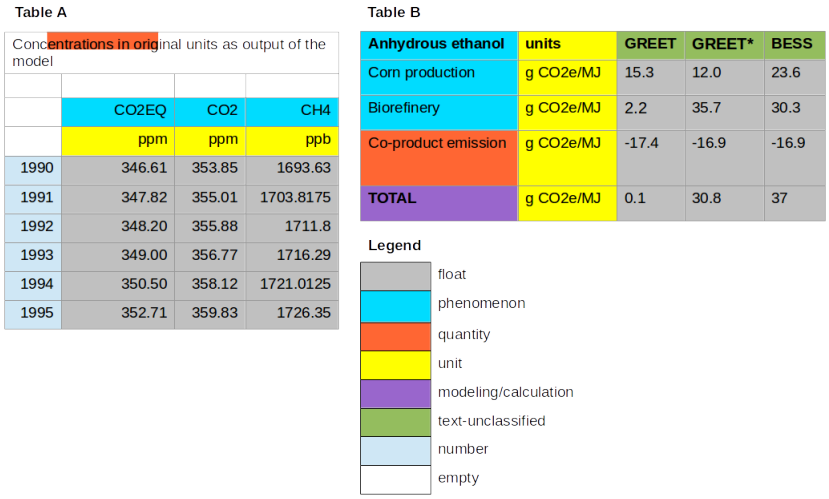

Table A

| Concentrations in original units as output of the model | | | |
|---|---|---|---|
| | CO2EQ | CO2 | CH4 |
| | ppm | ppm | ppb |
| 1990 | 346.61 | 353.85 | 1693.63 |
| 1991 | 347.82 | 355.01 | 1703.8175 |
| 1992 | 348.20 | 355.88 | 1711.8 |
| 1993 | 349.00 | 356.77 | 1716.29 |
| 1994 | 350.50 | 358.12 | 1721.0125 |
| 1995 | 352.71 | 359.83 | 1726.35 |

Table B

| Anhydrous ethanol | units | GREET | GREET* | BESS |
|---|---|---|---|---|
| Corn production | g CO2e/MJ | 15.3 | 12.0 | 23.6 |
| Biorefinery | g CO2e/MJ | 2.2 | 35.7 | 30.3 |
| Co-product emission | g CO2e/MJ | -17.4 | -16.9 | -16.9 |
| TOTAL | g CO2e/MJ | 0.1 | 30.8 | 37 |

Legend



**Figure 2: Examples of spreadsheet tables in which the quantity is present in the title (A),the units of measure are associated with phenomena cells (A,B), phenomenon cells contain abbreviations or numbers (A,B), the units of measure are customized (B). The color markup is applied in our analysis, and not part of the original table.**

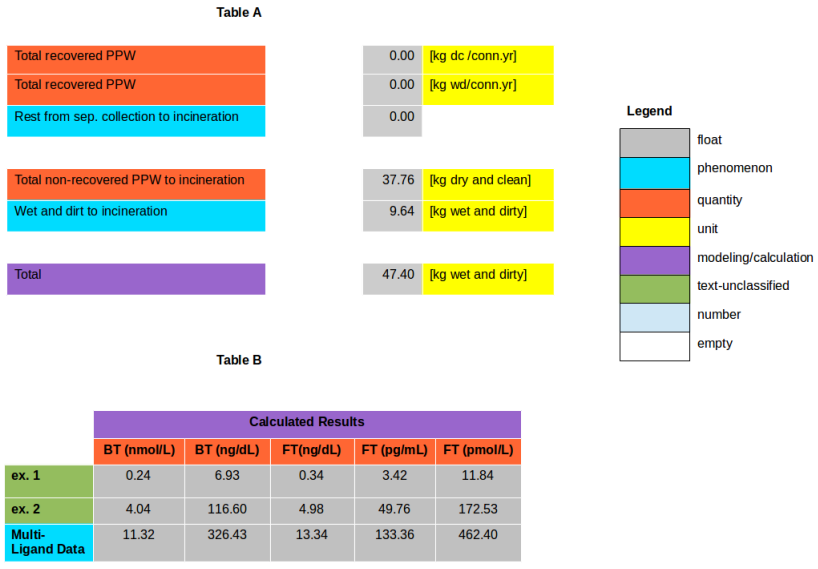

Table A

| | | |
|---|---|---|
| Total recovered PPW | 0.00 | [kg dc /conn.yr] |
| Total recovered PPW | 0.00 | [kg wd/conn.yr] |
| Rest from sep. collection to incineration | 0.00 | |
| Total non-recovered PPW to incineration | 37.76 | [kg dry and clean] |
| Wet and dirt to incineration | 9.64 | [kg wet and dirty] |
| Total | 47.40 | [kg wet and dirty] |

Legend



Table B

| | Calculated Results | | | | |
|---|---|---|---|---|---|
| | BT (nmol/L) | BT (ng/dL) | FT(ng/dL) | FT (pg/mL) | FT (pmol/L) |
| ex. 1 | 0.24 | 6.93 | 0.34 | 3.42 | 11.84 |
| ex. 2 | 4.04 | 116.60 | 4.98 | 49.76 | 172.53 |
| Multi-Ligand Data | 11.32 | 326.43 | 13.34 | 133.36 | 462.40 |

**Figure 3: Example of spreadsheet tables with interrupted float and string blocks (A), customized units of measure (A), and phenomenon cells with codes (B). The color markup is applied in our analysis, and not part of the original table.**

- *A block is considered a "Quantity" or a "Unit" block when at least 30% of the cells is recognized as a quantity or unit cell*
- *A block is considered a "Quantity" block when it is vertically or horizontally aligned with the "Unit" block and the float block*

Although the quantity cells in Figure 2B and 3B contain no description of a quantity concept, these cells could still be recognized as quantity cells by considering the presence and position of the units of measure in the table.

## 5.2 Knowledge from vocabularies

The phenomenon cells in natural science tables may not contain explicit terms, but codes, numbers or abbreviations (Section 4.1),

that are commonly used by scientists to refer to domain specific entities, e.g., chemical elements or human hormones (resp. Figure 2A and 3B). Vocabularies with additional information at the instance level could be used to annotate the content of these cells, and to recognize these as phenomenon cells. Furthermore, these vocabularies may also facilitate recognition of these phenomenon cells, by distinguishing the codes and abbreviations from quantities and units of measure.

Many quantity cells in the analyzed tables do not contain a concept description, but do have an associated unit of measure (Section 4.1). The missing quantity concept may be obtained by using the following heuristic:

- *Annotation concepts for quantity cells can be deduced from the included unit term*

Some units are commonly associated with certain quantities [23], and this type of information is included in, for example, the OM vocabulary. For example, the missing quantity concept in a cell containing the term "BT (nmol/L)" (Figure 3B) is probably "Molar Concentration", as the unit of measure "mol/L" is commonly associated with this quantity. However,this requires both a correct notation and interpretation of the associated units of measure, and that the information on the 'common association' is indeed present in OM.

Besides, domain scientists often customize the units of measure in their spreadsheets by combining unit symbols with phenomenon terms. Recognizing these phenomena using a domain vocabulary, and subsequently removing these from the unit terms would probably result in better recognition and interpretation of the units of measure in a table.

### 5.3 Deduction from table context

Empty cells in tables are often left empty on purpose, as data on either the observation value or its context is missing. In many cases these empty cells are surrounded by non-empty neighbouring cells, which could be used to deduce the missing information :

- *The content type of an empty cell is the same as that of the neighbouring cells*

Annotation of a whole group of phenomena is often difficult, as the semantic relation between the grouped phenomenon instances can not be recognized in a domain vocabulary. As suggested by [1, 3, 10], the following heuristic may be used:

- *A group of phenomenon instances may have a common denominator cell that is present above or left from the phenomenon block*

The term in this common denominator cell may be annotated and provide the concept of the phenomenon class.

## 6 PLAUSIBILITY OF AUTOMATIC ANNOTATION

In this section we investigate the applicability of our heuristics on the set of tables analyzed in our survey. Given that for the majority of these tables automatic annotation solely based on lexical matching would not be successful, the applicability of the heuristics gives us as an indication to what extent automatic annotation may still be possible. We distinguish three levels of plausibility:

*Automatic annotation is not possible.* A small part of the analyzed tables display serious deviations in their basic structure (e.g., Figure 3A). In these tables the blocks with numerical data on observations either are accompanied by only one block of contextual information, or from the table structure it is not clear which string and float cells are related to each other.

Although a good basic structure is not a requirement for successful lexical matching, it is a prerequisite for our heuristics. Without the presence and alignment of string and float blocks, our block heuristics (Section 5.1) can not be used to recognize the units of measure, quantities and phenomena in these tables. Consequently, without knowledge of the types of cells, deduction of information from the table context (Section 5.3) and deriving additional knowledge from vocabularies (Section 5.2) is not possible. The annotation of these tables would thus be based solely on lexical matching. As these tables do not contain complete and explicit information on the underlying research, we expect that automatic annotation of these tables is not possible. By the way, the majority of these tables would probably be hard to interpret for human readers as well.

*Automatic annotation is difficult.* Several of the analyzed tables have a good basic structure, but are missing entities.

In some of these tables the units of measure are missing, which hinders the recognition and annotation of quantities. The recognition of quantity cells in a table may be improved by block heuristics (Section 5.1). For the annotation of these quantities it is, however, not possible to derive additional knowledge from vocabularies.

In other tables the quantities are only implicitly represented (e.g., Figure 2A). Block heuristics may facilitate recognizing which of the contextual blocks in the table serves as a quantity block. The annotation of the quantities in these tables is difficult, as these are technically not present, but may be deduced from the associated units of measure.

*Reconstruction is possible.* The majority of the analyzed tables have a good basic structure, and consist of unit, quantity, and phenomenon cells, but do not contain complete and explicit information on these entities. As we expect all of our heuristics to be applicable in these type of tables, the missing information may be complemented and succesful recognition and annotation of the entities in these tables is possible.

## 7 DISCUSSION AND CONCLUSION

In this study we show that it is plausible to automatically annotate natural science spreadsheets, even if the tables do not contain complete and explicit information on the corresponding research project.

The quality and the level of detail of the annotations will, of course, still be depending on the completeness and accuracy of the content of the tables. However, even if the quality of the content is not sufficient to automatically annotate terms on an individual level, the block heuristics may be used to recognize the quantities, phenomena and units of measure blocks, thereby providing a basic understanding of the table. What is more, we think that block heuristics are useful in all spreadsheet tables, as these heuristics provide insight in how the cells in a table are related, and as such facilitate interpretation.

The number of existing research spreadsheets, especially in the informal area, is large. Our proposed approach could be used to facilitate search, reuse and integration of these spreadsheet data, by analyzing the information in annotations. Furthermore, the observations and heuristics from this study can be used as guidelines or in support tools for domain scientists to design new spreadsheet tables that are easier to interpret by both humans and machines. However, we do not believe that the common practice of spreadsheet development by domain scientists is easily changed. We expect that domain scientists will keep using spreadsheets, as these structures provide an easy and accessible way to store and manipulate reseach data according to their preferences. Therefore we expect that the automatic annotation of existing natural science spreadsheets will remain an issue. Our proposed method provides part of the solution to handle this issue.

## ACKNOWLEDGMENTS

This publication was supported by Dutch national program COMMIT.

## REFERENCES


[1] Robin Abraham and Martin Erwig. 2006. Inferring Templates from Spreadsheets. In *Proceedings of the 28th international conference on Software engineering*. ACM, 182–191.

[2] Michael J Cafarella, Alon Halevy, Daisy Zhe Wang, Eugene Wu, and Yang Zhang. 2008. WebTables: Exploring the Power of Tables on theWeb. *Proceedings of the VLDB Endowment* 1, 1 (2008), 538–549. https://doi.org/10.14778/1453856.1453916

[3] Zhe Chen and Michael J Cafarella. 2013. Automatic web spreadsheet data extraction. In *Proceedings of the 3rd International Workshop on Semantic Search Over the Web - SS@ '13*. 1–8. https://doi.org/10.1145/2509908.2509909

[4] Martin J O Connor, Christian Halaschek-wiener, and Mark A Musen. 2010. Mapping Master : a Flexible Approach for Mapping Spreadsheets to OWL. In *The Semantic WebâĂŞISWC*. Springer Berlin Heidelberg, 194–208.

[5] Martine G De Vos, Willem Robert Van Hage, Jan Ros, and Guus Schreiber. 2012. Reconstructing Semantics of Scientific Models : a Case Study. In *Proceedings of the OEDW workshop on Ontology engineering in a data driven world, EKAW 2012*. Galway, Ireland.

[6] Martine G De Vos, Jan Wielemaker, Hajo Rijgersberg, Guus Schreiber, Bob Wielinga, and Jan Top. 2017. Combining Information on Structure and Content to Automatically Annotate Natural Science Spreadsheets. *International Journal of Human-Computer Studies* (in press), 0 (2017).

[7] Martine G De Vos, Jan Wielemaker, Bob Wielinga, Guus Schreiber, and Jan Top. 2015. A methodology for constructing the calculation model of scientific spreadsheets. In *Proceedings of the 8th International Conference on Knowledge Capture*.

[8] Andres Garcia-silva, Asuncion Gomez-perez, Mari Carmen Suarez-figueroa, and Boris Villazon-terrazas. 2008. A Pattern Based Approach for Re-engineering Non-Ontological Resources into Ontologies. In *The Semantic Web*. Number 2. Springer Berlin Heidelberg, 167–181.

[9] Lushan Han, Tim Finin, Cynthia Parr, Joel Sachs, and Anupam Joshi. 2008. RDF123 : From Spreadsheets to RDF. In *The Semantic Web-ISWC 2008*. Springer Berlin Heidelberg, 451–466.

[10] Felienne Hermans, Martin Pinzger, and Arie Van Deursen. 2010. Automatically Extracting Class Diagrams from Spreadsheets. In *24th European Conference on Object-Oriented Programming (ECOOP),Lecture Notes in Computer Science,*. Springer-Verlag, 52–75.

[11] Andreas Langegger and W Wolfram. 2009. XLWrap âĂŞ Querying and Integrating Arbitrary Spreadsheets with SPARQL. In *International Semantic Web Conference*. 359–374.

[12] Girija Limaye, Sunita Sarawagi, and Soumen Chakrabarti. 2010. Annotating and searching web tables using entities, types and relationships. In *Proceedings of the VLDB Endowment*, Vol. 3. 1338–1347. https://doi.org/10.14778/1920841.1921005

[13] Eamonn Maguire, Alejandra González-Beltrán, Patricia L. Whetzel, Susanna Assunta Sansone, and Philippe Rocca-Serra. 2013. OntoMaton: A Bioportal powered ontology widget for Google Spreadsheets. *Bioinformatics* 29, 4 (2013), 525–527. https://doi.org/10.1093/bioinformatics/bts718

[14] Daniel McDonald, Jose C Clemente, Justin Kuczynski, Jai Rideout, Jesse Stombaugh, Doug Wendel, Andreas Wilke, Susan Huse, John Hufnagle, Folker Meyer, Rob Knight, and J Caporaso. 2012. The Biological Observation Matrix (BIOM) format or: how I learned to stop worrying and love the ome-ome. *GigaScience* 1, 1 (2012), 7. https://doi.org/10.1186/2047-217X-1-7

[15] Albert Meroño-Peñuela, Ashkan Ashkpour, Laurens Rietveld, Rinke Hoekstra, and Stefan Schlobach. 2013. Linked Humanities Data : The Next Frontier ? A Case-study in Historical Census Data. In *The Semantic Web: Semantics and Big Data*. Springer Berlin Heidelberg, 645–649.

[16] Roland T. Mittermeir and Markus Clermont. 2002. Finding High-Level Structures in Spreadsheet Programs. In *Proceedings of the 9th Working Conference on Reverse Engineering*. Richmond,VA,USA, 221–232.

[17] Varish Mulwad, Tim Finin, and Anupam Joshi. 2012. A Domain Independent Framework for Extracting Linked Semantic Data from Tables. In *Search Computing*. Springer Berlin Heidelberg, 16–33.

[18] Tim F Rayner, Philippe Rocca-Serra, Paul T Spellman, Helen C Causton, Anna Farne, Ele Holloway, Rafael A Irizarry, Junmin Liu, Donald S Maier, Michael Miller, Kjell Petersen, John Quackenbush, Gavin Sherlock, Christian J Stoeckert, Joseph White, Patricia L. Whetzel, Farrell Wymore, Helen Parkinson, Ugis Sarkans, Catherine A Ball, and Alvis Brazma. 2006. A simple spreadsheet-based, MIAME-supportive format for microarray data: MAGE-TAB. *BMC bioinformatics* 7 (2006), 489. https://doi.org/10.1186/1471-2105-7-489

[19] Hajo Rijgersberg, M. Wigham, and Jan Top. 2011. How semantics can improve engineering processes: A case of units of measure and quantities. *Advanced Engineering Informatics* 25, 2 (apr 2011), 276–287. https://doi.org/10.1016/j.aei.2010.07.008

[20] Ivelize Rocha Bernardo, Matheus S Mota, and André Santanchè. 2013. Extracting and Semantically Integrating Implicit Schemas from Multiple Spreadsheets of Biology based on the Recognition of their Nature. *Journal of Information and Database Management* 4, 2 (2013), 104–113.

[21] Susanna-Assunta Sansone, Philippe Rocca-Serra, Dawn Field, Eamonn Maguire, Chris Taylor, Oliver Hofmann, Hong Fang, Steffen Neumann, Weida Tong, Linda Amaral-Zettler, Kimberly Begley, Tim Booth, Lydie Bougueleret, Gully Burns, Brad Chapman, Tim Clark, Lee-Ann Coleman, Jay Copeland, Sudeshna Das, Antoine de Daruvar, Paula de Matos, Ian Dix, Scott Edmunds, Chris T Evelo, Mark J Forster, Pascale Gaudet, Jack Gilbert, Carole Goble, Julian L Griffin, Daniel Jacob, Jos Kleinjans, Lee Harland, Kenneth Haug, Henning Hermjakob, Shannan J Ho Sui, Alain Laederach, Shaoguang Liang, Stephen Marshall, Annette McGrath, Emily Merrill, Dorothy Reilly, Magali Roux, Caroline E Shamu, Catherine A Shang, Christoph Steinbeck, Anne Trefethen, Bryn Williams-Jones, Katherine Wolstencroft, Ioannis Xenarios, and Winston Hide. 2012. Toward interoperable bioscience data. *Nature genetics* 44, 2 (2012), 121–6. https://doi.org/10.1038/ng.1054

[22] Yanfeng Shu, David Ratcliffe, Michael Compton, Geoffrey Squire, and Kerry Taylor. 2015. A semantic approach to data translation: A case study of environmental observations data. *Knowledge-Based Systems* 75 (2015), 104–123. https://doi.org/10.1016/j.knosys.2014.11.023

[23] Mark Van Assem, Hajo Rijgersberg, M. Wigham, and Jan Top. 2010. Converting and Annotating Quantitative Data. In *ISWC2010*, P.F. Patel-Schneider (Ed.). 16–31.

[24] Petros Venetis, Alon Halevy, and J Madhavan. 2011. Recovering semantics of tables on the web. In *Proceedings of the VLDB Endowment*, Vol. 4. 528–538. https://doi.org/10.14778/2002938.2002939

[25] Katy Wolstencroft, Stuart Owen, Matthew Horridge, Olga Krebs, Wolfgang Mueller, Jacky L Snoep, Franco du Preez, and Carole Goble. 2011. RightField: embedding ontology annotation in spreadsheets. *Bioinformatics (Oxford, England)* 27, 14 (jul 2011), 2021–2. https://doi.org/10.1093/bioinformatics/btr312