FOOD & BIOBASED RESEARCH
WAGENINGEN UR

COMMIT/

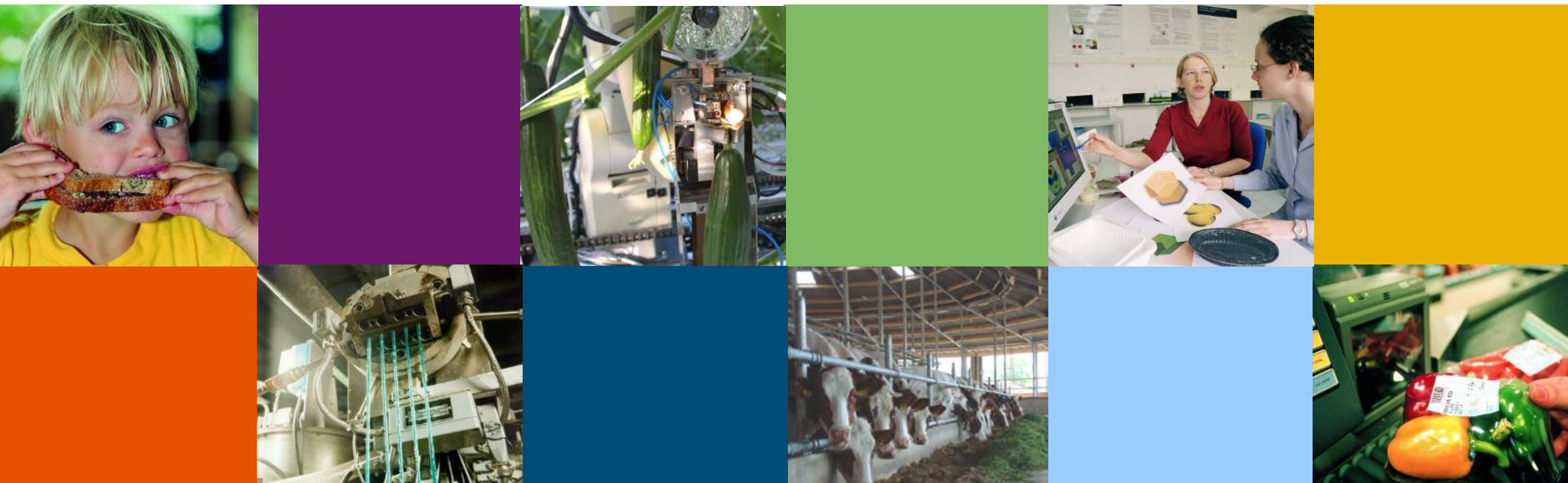
Rosanne: islands of structure in unstructured data

Final report 2015

Rosanne valorisation project

Mari Wigham
Hajo Rijgersberg
Mariëlle Timmer
Jan Top

Report 1590



Colophon

Title	Rosanne: islands of structure in unstructured data
Author(s)	M. Wigham, H.Rijgersberg, M. Timmer, J. Top
Number	1590
ISBN-number	978-94-6257-523-3
Date of publication	01-06-2015
Confidentiality	No
Approved by	Jan Top

Wageningen UR Food & Biobased Research
P.O. Box 17
NL-6700 AA Wageningen
Tel: +31 (0)317 480 084
E-mail: info.fbr@wur.nl
Internet: www.wur.nl

1 Introduction

Different industrial food companies and research centres have expressed a need for data management. In industrial research and development, but also in production settings, trade and retail, spreadsheets with unstructured data are abundant [1]. The terminology used in these datasets typically differs between individuals, departments and organizations. Moreover, jargon and personal abbreviations are often used; units of measure are omitted, etc. Lack of understanding of the data and needless overhead for clarifying the data causes work processes to be inefficient. It can also give rise to serious errors [2], [3], [4]. Without high quality metadata it is very difficult to find, interpret and reuse the data, let alone integrate data from different sources [5].

Rosanne [6] is a software tool that aims to facilitate adding 'islands of structure' in terms of semantic tables within otherwise unstructured formats. This means that tables are annotated and handled using shared vocabularies. These annotations supply the necessary metadata to make it easy to find, interpret and reuse the data. Rosanne also offers support in integrating data from different spreadsheets, using the information in the annotations.

The aim of this valorisation project was to produce a proof-of-concept version of Rosanne, which would offer all functionality within Excel and be robust, fast and user-friendly. In addition, it would link to company ontologies, created in this project, to offer easy customisation to each company's unique needs while maintaining generic applicability. An essential part of the project was to test Rosanne with several industrial partners.

This report will first briefly describe Rosanne in Section 2. Section 3 will discuss the industrial use cases and the test findings. Section 4 will briefly list the improvements made to Rosanne during this project, while Section 5 will lay out the necessary areas for future improvement. Section 6 will describe the dissemination activities. Finally, Section 7 will conclude with a discussion of the project results in relation to the plan.

2 Rosanne

Rosanne[6] is a software tool that allows users to add semantics when creating new datasets. It is implemented as an add-in for the popular Microsoft Excel software package. Users select areas inside a sheet to become semantic tables. They can then select terms from proprietary or public ontologies, including the publicly available and widely applicable OM-ontology [7] (quantities and units of measure), to annotate the data in the table.

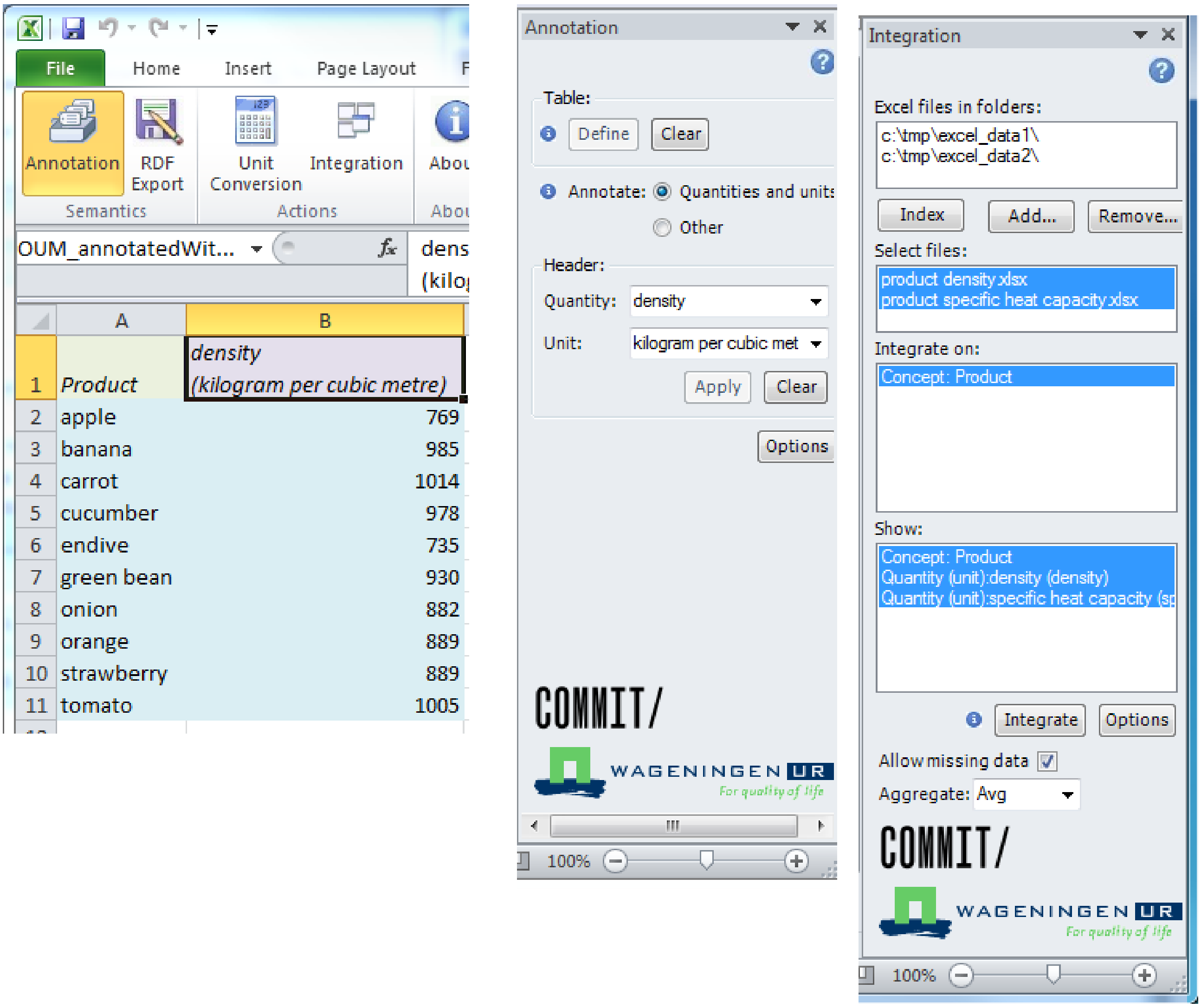
These vocabularies (ontologies [8]) ensure that unambiguous concepts are used, making it easier to find and understand the data. However, users still have the freedom to use their own text in the headers (for example departmental terminology, or a local language) and other areas of the spreadsheet.

Behind the screens, Rosanne uses the RDF Record Table model [6] to model the structure of the annotated data table. This permits the data to be converted to an entirely semantic form, which enables advanced computer support in processing the data. Rosanne takes advantage of this to offer users support when integrating files. Users can select two or more spreadsheet files, and select any number of tables from the spreadsheets for integration. The user does not have to know the cell address of the data or even which file it is in. They define the integration easily in terms of the semantic concepts; for example they may choose to see 'mass', 'viscosity' and 'creaminess' of a 'Product'. The add-in then compiles the table with this information automatically.

Rosanne is seamlessly embedded in Excel, extending the standard ribbon to enable user actions. Once a user has obtained permission to install Rosanne, it can simply be downloaded and installed locally as a component of Excel. It connects to public ontologies using web services.

Rosanne is unique due to (i) its online connection to a library of ontologies and (ii) its easy user interface [6]. The ontology library includes the comprehensive ontology of quantities and units of measure OM [11] and several food-related ontologies, all available online on the website http://www.wurvoc.org. A user or company may also add their own ontologies to accurately describe their own data. Rosanne allows users to manipulate tables in terms of objects and quantities rather than indices of rows and columns; it is seamlessly integrated in standard spreadsheet software – essential for usability and user acceptance. Moreover, to the best of our knowledge, it is the only solution that has built-in support for semantic data integration, so that a direct benefit can be derived from the annotations in making integration quick and easy.

Figure 1 shows the annotation and integration functionality of Rosanne in action.



	A	B
1	*Product*	*density (kilogram per cubic metre)*
2	apple	769
3	banana	985
4	carrot	1014
5	cucumber	978
6	endive	735
7	green bean	930
8	onion	882
9	orange	889
10	strawberry	889
11	tomato	1005

Table: Define Clear

Annotate: ● Quantities and units ○ Other

Header:

Quantity: density

Unit: kilogram per cubic met

Apply Clear

Options

Excel files in folders:
c:\tmp\excel_data1\
c:\tmp\excel_data2\

Index Add... Remove...

Select files:
product density.xlsx
product specific heat capacity.xlsx

Integrate on:
Concept: Product

Show:
Concept: Product
Quantity (unit):density (density)
Quantity (unit):specific heat capacity (sp

Integrate Options

Allow missing data ☑

Aggregate: Avg

Figure 1. Sections of the Microsoft Excel add-in Rosanne. Left is a simple table example. The annotation pane is shown in the centre, displaying the annotations contained in the selected cell. The integration pane is shown on the right. The user has selected the field 'Product' as the field on which the tables will be joined. In the integrated table she also wants to include the quantities 'density' and 'specific heat capacity'.

3 Testing and evaluation on industrial use cases

We requested use cases for annotation and/or integration of data from five companies; Corbion[1], Danone[2], Friesland Campina[3], Unilever[4] and Wageningen UR (Food and Biobased Research[5]). We received one or more use cases each from four of the five companies; the fifth company did not respond with a use case within the timeframe of the project. As these cases can involve confidential research data, we will not discuss details of the content or analysis. The use cases consisted of multiple Excel spreadsheets, together with a specification of the desired output.

The original plan was to allow users from the companies to carry out the integration themselves, after a brief introduction to the Rosanne tool. In the event, the data in the use cases turned out to be far more complex in its structure than expected. We had anticipated that the simple structure supported by Rosanne would need to be expanded to allow nesting in tables (which is already supported by the underlying RDF Record Table model). However, this was not in fact the type of complexity which occurred in the files. Data was far more fragmented and implicit than expected; for example data was split over multiple tables in one sheet, variables were omitted as they were implicit in the filename, and empty fields were sometimes used to convey the information 'same as the above'. Figure 2 shows dummy data that demonstrates these issues.

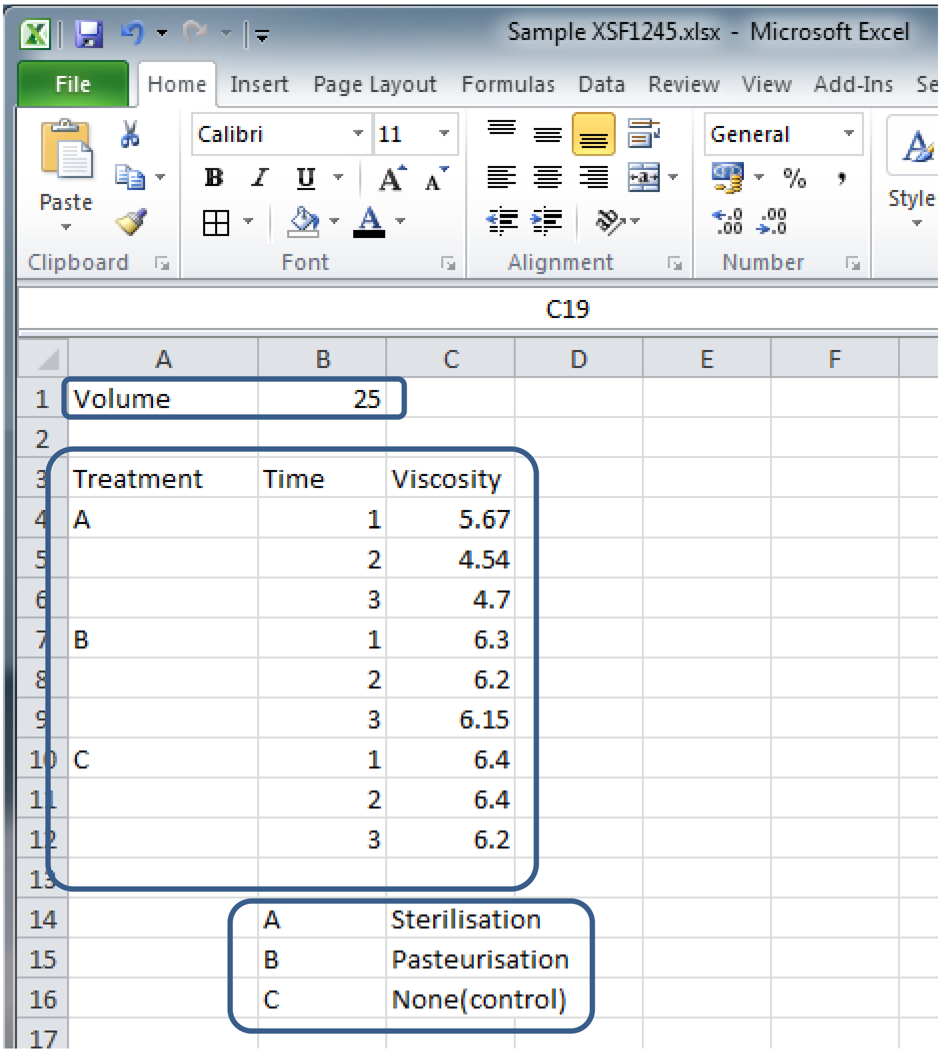
[1] *http://www.corbion.com/*

[2] *http://www.danone.nl/*

[3] *http://www.frieslandcampina.com/english*

[4] *http://www.unilever.nl/*

[5] *http://www.wageningenur.nl/en.htm*



	A	B	C	D	E	F
1	Volume	25				
2						
3	Treatment	Time	Viscosity			
4	A	1	5.67			
5		2	4.54			
6		3	4.7			
7	B	1	6.3			
8		2	6.2			
9		3	6.15			
10	C	1	6.4			
11		2	6.4			
12		3	6.2			
13						
14		A	Sterilisation			
15		B	Pasteurisation			
16		C	None(control)			
17						

Figure 2: Dummy data demonstrating some of the data complexity issues. This sheet contains data split over three different tables. Two of the tables refer to the sample XSF1245, but this sample is only mentioned in the file name. Rows 4, 5 and 6 all refer to treatment A, however this is only explicitly entered in Row 4.

As a result of this, several manual steps were required to clean up the data to the point at which it could be annotated and integrated using Rosanne, and to carry out the integration fully.
For example, tables had to be copied to other sheets, variables had to be inserted and empty cells filled with the correctly interpreted data. Integrations involving more than two tables also had to be performed as multiple pair-wise integrations. In the course of the project, we improved the functionality of Rosanne to make a number of these steps unnecessary (see Chapter 4).
However, the time available was insufficient to develop Rosanne enough to eliminate all the manual steps. We therefore carried out the clean-up, annotation and integration process ourselves, and gave the result to the users for evaluation.

Two use cases turned out to actually involve regression, rather than integration. Integration involves merging (a selected subset of) the information in multiple tables into one table. Regression is a statistical process for estimating the relationships between variables, and as such is an analysis of the table data, not an integration. For the remaining cases we were able to successfully annotate the data and then carry out the desired integration using Rosanne, after manual clean-up.

Due to the necessity of the manual steps discussed above, it was not possible to carry out the planned comparison of the time required to locate and integrate data with and without Rosanne. This comparison must be rescheduled to be carried out once these manual steps have been eliminated (as will be discussed in Section 5).

During this project it became clear that the introduction of good quantitative data annotation represents a large cultural shift for companies. While the industrial partners in this project all had strong advocates for good data documentation, the necessity and the benefits were not yet broadly understood within the company. The introduction of Rosanne was therefore awkward, as there was not yet sufficient awareness of the problem it was intended to solve. Initially, researchers were very unclear as to what Rosanne was for, which use cases could be relevant, and what the benefits could be. This slowed the process of gathering and implementing the use cases considerably.

Once a use case had been completed and the results presented back to the users, they understood the benefits, and this led to an enthusiastic discussion of where else they could use Rosanne, and requests for more tests and new features. For some of the use cases there was a clear wish to extend this sort of semantic support to earlier points in the data process, such as experimental design, and later points, such as statistical analysis.

We had expected to construct domain ontologies together with the industrial participants, to allow Rosanne to offer annotations tailored to the work area of each company. As we did not reach the stage of allowing users to test Rosanne themselves, there was also not yet the necessity to build a domain ontology. At the same time, the necessary level of involvement from the industrial partners for building an ontology was not yet present, due to the need to first create awareness and understanding, as described above.

.

4 Improvements made to Rosanne

As a result of the feedback given during the use cases, we have made a number of significant improvements to Rosanne. These are discussed in detail in the eFoodLab deliverables [9] and [10], and as such will be briefly summarised here

- Annotating individual instances with unique identifiers – Not only can a column be annotated with, e.g. ‘Cheese’, but the individual items in the column can be annotated with unique identifiers, e.g. ‘AH Gouda jong belegen’.
- Annotating more than one table per sheet – Each Excel spreadsheet can contain any number of tables
- Integrating more than two files - Any number of files can be integrated together in one single step
- Integrating more than two tables – Any number of tables can be integrated together in one go
- Missing values – Rosanne can handle missing values in the data during integration
- Automatic annotation after integration – Integration results are automatically annotated
- Original text – The original text in the headers is included in the integration result

These improvements make Rosanne much more practical and easy to use. In particular the ability to create multiple tables in each sheet and to integrate any number of tables in one single step represents an important advance.

5 Issues for further development

We identified two types of issues that need to be tackled in order to make Rosanne into a viable product: design issues, which need to be solved by changing the design of Rosanne, and support issues, which need to be addressed by increased user support. These issues are described in detail in the eFoodLab deliverable [10] and as such will be briefly summarised here.

The most important design issues are:

- Handling complex or unusual table structures – Rosanne needs to offer more flexibility in the allowed structure of a table
- Joining on unique identifiers during integration – Matching records for integration was done by string matching, this needs to be done using the unique identifiers in the annotations, otherwise minor typing errors can block integration
- More advanced statistical analysis – Users would like to carry out analysis, such as correlation. This could be done by linking Rosanne to statistical software
- Provenance – it is important to be able to track where the original data in an integrated file came from
- Performance – Rosanne needs to be faster at processing large files. While small tables are integrated in a couple of seconds, the largest files we have tested, with thousands of records, take several hours to integrate.
- Explanation of options – Rosanne offers several options when integrating data, the effect of these options on the result needs to be clearer

Note that the first issue, that of handling complex or unusual table structures, can be tackled partly by improving Rosanne, and partly by better awareness among spreadsheet users as to how to record their data. This brings us on to the discussion of support issues.

It became clear during this project that the issue of data documentation is not widely appreciated. As a result, users need support to help them to understand the potential benefits, and to identify how they can improve their data processes. An essential part of this is to think not only in terms of what they need themselves from their data, but also in terms of what their (future) colleagues need. This leads to more transparent, reusable data. This support is therefore an integral part of the introduction of Rosanne, or similar software, to a company. The company needs to be supported through a process of cultural change in order to reach a higher 'maturity level' in how they handle recording and sharing data.

Specific plans for tackling both design and support issues are laid out in the eFoodLab deliverable [10].

Note that our work in the related COMMIT/eFoodLab project, WP2, has also identified other areas of future development for Rosanne, such as automatic data annotation, central databases, and automatic unit conversion. While these would of course also benefit the industrial users from this project, we will not discuss them here as they are not directly relevant to the use cases. Please refer to the eFoodLab deliverables for further information.

6 Dissemination

The knowledge generated during this project has been disseminated via regular deliverables, which have been made available through the COMMIT/ eFoodLab project. This final report will also be published as a COMMIT/ deliverable, and will be distributed to all industrial participants.

Rosanne has been presented at the board meeting of the TI Food and Nutrition, a research alliance of 26 companies and research centres. A short film describing the concepts behind Rosanne is available on the Food and Biobased Research YouTube channel [11].

Rosanne and its integration functionality were published in an extended journal paper on the underlying RDF Record Table [12]. We also plan to submit an article about Rosanne to the December 'Trends en toekomst' issue of the VMT magazine for food professionals.

7 Conclusion

The first conclusion is that Rosanne is at an earlier stage in the valorisation process than was apparent in tests within research institutes. The industrial data and use cases demonstrate a pressing need for greater flexibility. We have identified possible solutions by which we can build this greater flexibility into Rosanne. The feedback obtained from industrial users will be invaluable in shaping Rosanne into a tool that will enable them to get more benefit out of their spreadsheet data.

The second conclusion is that Rosanne represents, for most companies, not just a software tool but also a cultural shift. The industrial partners are at the forefront of this new development, as they are already aware of the pressing need for a change in how data is recorded, shared and reused. This awareness needs to spread from the strategic thinkers throughout the whole company to all the employees. Such a cultural shift cannot be realised by simply making a software tool available. It is essential to embed Rosanne in a broader context which includes working together with the company to raise awareness and to develop working procedures for semantic data.

This leads to the final conclusion that introduction of Rosanne in a company should be in the form of a workshop, in which a case study from the company is used as preparation. This will provide the workshop attendees with a concrete example of the benefits and act as an inspiration to help identify how Rosanne can be implemented in the company's processes. Any requests for extra functionality can then also be considered. Rosanne should then be gradually introduced, starting with small, enthusiastic teams, and later extending to include more users. This process should include regular opportunities for feedback and discussion between the company and the researchers at Food & Biobased Research.

This valorisation project has been a vital step in beginning this necessary partnership, by bringing together industrial users and experts in semantic support for data. The potential users have shown interest not only in Rosanne for annotation and integration, but also in using the annotation techniques earlier in the research workflow to capture more information about experimental design, and in using the information in the annotations to support statistical analysis. This opens up possibilities for semantic support throughout the workflow. This process will be continued within the COMMIT/ eFoodLab project and in future bilateral projects with the industrial partners.

8 References

[1] Wayne L. Winston. "Executive Education Opportunities," OR/MS Today, August (2001).

[2] 'Spreadsheet blunders costing business billions - CNBC.com". http://www.cnbc.com/id/100923538 Retrieved 3rd July 2015

[3] C. S. Jensen and R. T. Snodgrass, "Temporal Data Management," IEEE Transactions on Knowledge and Data Engineering, vol. 11, January/February 1999, pp. 36-44.

[4] O. Etzion, S. Jajodia, and S. Sripada (eds.), "TemporalDatabases: Research and Practice," LNCS 1399, Springer-Verlag, 1998.

[5] C. S. Jensen and R. T. Snodgrass, "Temporal Database," inLiu L., Özsu M.T., (Eds.), Encyclopedia of Database Systems, Springer US, 2009, pp. 2957-2960.

[6] J. Top, H. Rijgersberg, M. Wigham, "Semantically enriched spreadsheet tables in science and engineering", Eighth International Conference on Advanced in Semantic Processing, Rome, Italy, 2014.

[7] H. Rijgersberg, M. Wigham, and J. L. Top, "How semantics can improve engineering processes: A case of units of measure and quantities.""Advanced Engineering Informatics, 25(2), 2010, pp.276–287.

[8] Gruber, Thomas. "Toward Principles for the Design of Ontologies Used for Knowledge Sharing". International Journal Human-Computer Studies Vol. 43, Issues 5-6, November 1995, p.907-928.

[9] Rijgersberg et al., "Experiments with spreadsheet data in industrial R&D", Deliverable Q3 2014, eFoodLab WP2.

[10] Wigham et al., "Rosanne - specification of the Proof of Concept", Deliverable Q3 2014, eFoodLab WP2.

[11] Film, "Sharing your research with Tiffany and Rosanne" https://www.youtube.com/watch?v=fdFrJuf26r4

[12] Wigham et al., "Semantic Support for Tables using RDF Record Table", International Journal On Advances in Intelligent Systems, v 8 n 1&2 2015.