FOOD & BIOBASED RESEARCH
WAGENINGEN UR

Semantic support for data analysis

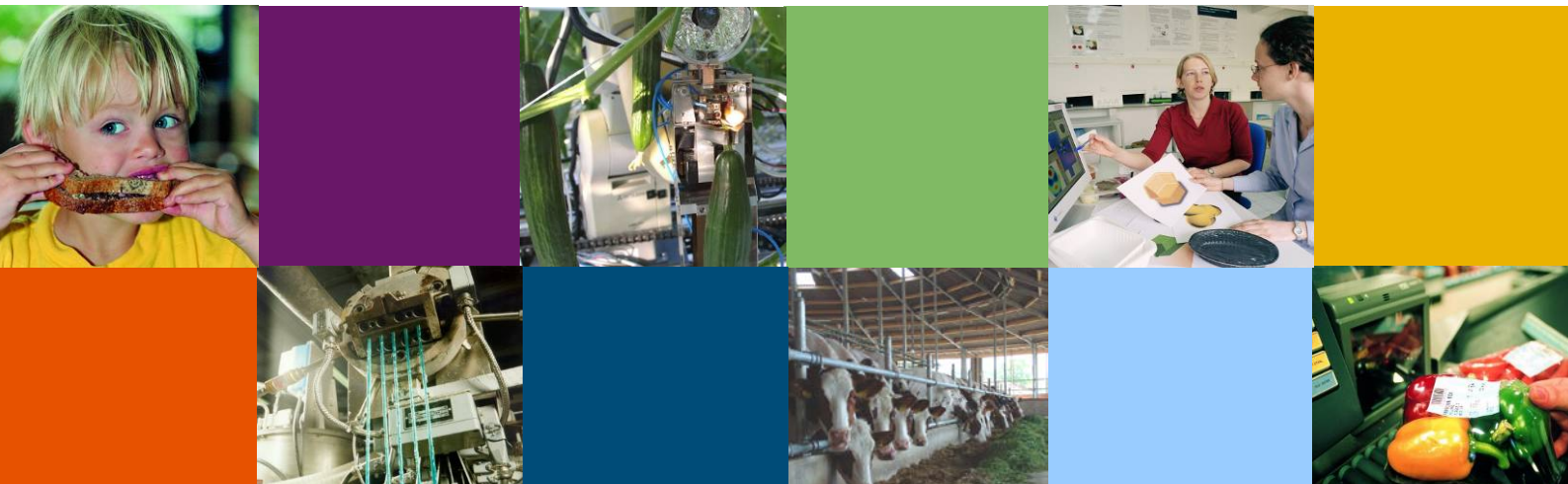
Deliverable Q2 2014

eFoodLab WP2

Eric Boer, Don Willems, Jan Top

COMMIT/

Rapport nr. 1524



Colophon

Title	Semantic support for data analysis
Author(s)	E. Boer, D. Willems, J. Top
Date of publication	15-07-2014
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

I. Introduction

Semantic Web technologies enable people to create data stores on the Web, build vocabularies, and write rules for handling data [1]. This document will explore how semantic support can assist researchers with their data analysis. The complete experiment can be seen as a workflow, starting with a specific research idea and ending with a report including the data analysis, see Figure 1. We aim to develop semantic support for this entire workflow based on building blocks that can be reused in all kind of quantitative research workflows. We will demonstrate the principles by a use case in the field of food sensory science.

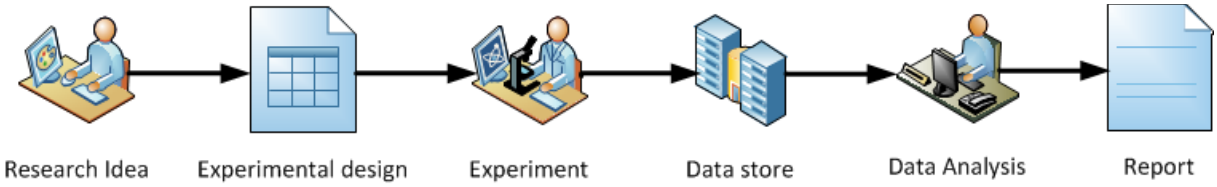


Figure 1 Workflow of quantitative research

Although the data analysis will be often in one of the final steps of the quantitative research workflow, every statistician will stress that it is extremely important to design the experiment in a statistical sound manner. A good experimental design should guarantee that the research questions can be addressed. The famous statistician Sir R.A. Fisher said at a congress in 1938: "To consult the statistician after an experiment is finished is often merely to ask him to conduct a post mortem examination. He can perhaps say what the experiment died of." The problem of improperly designed experiments cannot be solved by advanced statistical analysis. However, there are many scientists who try to disguise badly constructed experiments by blinding their readers with a complex statistical analysis [2].

Choices made at the stage of the experimental design have consequences for the data analysis. For example, if we would like to know the monthly salary of consumers in a survey, we could ask whether the individual consumer earns more or less than modal, or we could ask for their monthly salary in euros. The first question results in discrete (binary in this case) data, the second question in continuous data. If we would like to visualize this data, the discrete data should be presented by a bar diagram, the continuous data by a histogram. This simple example shows that choices during the design phase of the experiment lead to the application of different statistical methods.

Currently, in many studies the experimental design is separated from data analysis. This may be because multiple people are involved in the quantitative research workflow. Even if one person handles the complete workflow, the experimental design is often solely based on practical considerations and not on statistical reasons. Researchers often start to think about the data analysis at end of the quantitative workflow. Performing the data analysis without or with limited information about the experimental design is a major source of errors in quantitative research.

Semantic support in quantitative research is crucial for reuse and performing meta-analysis on the collected data. However, that is not the scope of this study. In this document, we demonstrate that semantic support can facilitate the workflow in quantitative research. The focus is on assisting the researcher in making the correct (statistical) choices in both experimental design and data analysis. In this study, the semantic support is given by data models for describing the experimental design and data. These data models annotate additional information to the experimental design and data, which can be used in the data analysis. These annotations can also be used for automation of data analysis.

II. Semantic support in quantitative research

In this paragraph we will explain how semantic support can be included in the quantitative workflow and how this can facilitate the research in several ways. Focus will be on semantic support for experimental design, data storage and data analysis.

1.1 Semantic support for experimental design

Practical considerations are often major factors during the design of an experiment. Designing the experiment in a statistical sound manner does not change these practical considerations. A more statistical mind-set can prevent important errors. When practical considerations are leading in the experimental design, there is a risk that the designed experiment cannot answer the original research question. For every research question specific choices should be made in the experimental design. For quantitative research, the experimental design has certain generic components. To name a few:

- The goal of every experiment should be formulated explicitly in terms of research questions.
- The population and experimental units should be defined. Which are the basic objects (a person, a sample of soil, a can of tomatoes, etc.) upon which the experiment is carried out?
- Define the variables that should be measured, including physical units and measuring scales of these variables.
- Which conditions in the experiment are changed systematically to answer the research questions? In statistical terms these conditions constitute the definition of the factors. How many levels are included in these factors? For example, a researcher could compare three different products for 'liking'. In that case, 'Kind of product' is called the factor and the different products (A, B and C) are the levels of that factor.
- Are there repeated measurements or not? How many measurements of one variable are done for each experimental unit? In statistical terms this means whether measurements are dependent or not. For example, when blood pressure of patients is measured before and after taking a drug, the two measurements of blood pressure are related (dependent), because they are measured on the same patient.

Semantic support can define these generic concepts in a data model. Given this data model the researcher can be assisted in systematically answering these questions. A logical choice is to build software in which these questions are handled. The answers to these questions are stored in a formal way using the data model.

1.2 Semantic support for data storage

Often, data of experiments are stored in spreadsheets like Excel. For a simple experiment the spreadsheet will contain an indicator of the experimental unit (name of a person, number of a sample), the different levels of a factor and the measurements of the variables, see Figure 2.

Sample	Factor	Variable 1	Variable 2	Variable 3
1	Level 1	2	3.34	Yes
2	Level 2	5	5.86	No
3	Level 3	6	7.89	Yes
4	Level 1	3	4.56	Yes
5	Level 2	7	4.75	No
6	Level 3	8	6.32	Yes
7	Level 1	3	3.01	No

Figure 2 Example of data storage in a spreadsheet like Excel

These spreadsheets contain the basic information for the data analysis, but a lot of information of the experiment is not expressed in this spreadsheet. Examples of hidden information are the research questions, the measurement scales of the variables or the exact questions in a survey (which question is answered for Variable 3). When the total workflow is performed by different people, all this additional information should be communicated between the different people involved in the workflow. In practice this is a major source of errors in the data analysis and can lead to wrong conclusions of the research. With use of data models to create a representation of the experimental design the data can be enriched with information. In this way, the data resulting from the experiment will be connected to the experimental design expressed in the same data model.

1.3 Semantic support for data analysis

Choices in the design of the experiment may lead to the application of different statistical methods. In the introduction the example is given of discrete and continuous data. For discrete data a bar chart is a logical graphical representation, and for continuous data a histogram or boxplot could be chosen. When the measurement scale of a variable is known, the researcher could be guided in the right direction. In many situations it will be clear that for discrete data the visualization should be a bar diagram. For continuous data you could ask the researcher, via a user interface, whether it is preferred to have a boxplot or a histogram.
Another example of semantic support relates to interdependency between measurements. In statistical analysis independence of measurements is very important. It should be known whether measurements are related or independent. If the experimental design is such that that there are related measurements, many statistical methods will fail. A researcher with less statistical knowledge can easily choose the wrong statistical method. Information given by the semantic

support can prevent this error by reminding the researcher for related measurements or give guidance in the choice of the correct statistical method.
The correct statistical method is often already defined by the experimental design. For routinely performed research or specific situations, the statistical analysis could be automated based on the information of the experimental design. When a complete automation of the data analysis is not possible or desired, the researcher could be guided to the correct choice of a statistical method.

III. Methodology

In our approach we express the experimental design and data in the Resource Description Framework (RDF) as specified by W3C [3]. RDF uses triples to link resources. Each triple consists of a subject, a predicate (usually expressed as a verb) and an object. Triples are the smallest unit of knowledge. For example, the sentence 'Oslo is the capital of Norway', consists of a subject 'Oslo', a predicate 'is the capital of' and an object 'Norway'. RDF models are often visualised as graphs, where subjects and objects are represented by nodes and the predicates are represented by the edges. One node (e.g. a subject) may be connected to many other nodes (objects). Another triple may for instance express 'Oslo (subject) was established (predicate) in 1048'. Using this simple triple-model, RDF allows structured and semi-structured data to be mixed, exposed, and shared across different applications [3].
To enable the reuse of resources, RDF uses URIs (Uniform Resource Identifiers of which URLs are a subtype) to uniquely identify resources. The URI 'http://sws.geonames.org/3143244/', for instance, uniquely identifies the city of Oslo in Norway in the Geonames dataset. Other datasets or applications may reuse this URI to identify Oslo, which enables the reuse and combination of data from different sources.
Sets of concepts that are important in a certain domain are often collected in ontologies. Ontologies define the types, properties and interrelations of concepts within an ontology or even linked to other ontologies. Using technologies related to RDF such as RDF Schema language (RDFS) and the Web Ontology Language (OWL), ontologies may be represented in RDF. Each concept within an ontology is uniquely defined by a URI, which can then be used to disambiguate knowledge across data sets or applications.
In our domain model we reuse existing ontologies such as the Ontology of Units of Measure (OM) which defines units and quantities used in quantitative research[13], PROV-O [4], which provides concepts used to express provenance information, and the Ontology for Quantitative Research (OQR) that defines the concepts and properties needed to express tabular data.
Both the experimental design and the data resulting from an experiment can be expressed using concepts from these ontologies. The data expressed in RDF and stored as RDF can be queried using the SPARQL query language [5].
In this study we use the statistical package R [6] to automate some statistical analysis, but this could be done in almost any other statistical package. R is an increasingly popular open source statistical package. R is considered to be a language and environment for a wide variety of statistical analysis (linear and nonlinear modelling, classical statistical tests, time-series analysis,

classification, clustering, ...) and graphics. The standard (or base) packages are considered part of the R source code. They contain the basic functions that allow R to work.
There are thousands of contributed packages for R, written by many different authors. Some of these packages implement specialized statistical methods, others give access to data or hardware, and others are designed to complement textbooks. In this document the package SPARQL[7] is used. This function connects to a SPARQL end-point over HTTP or HTTPs and returns the results as a data frame within R.

2 Case study: sensory evaluation of cappuccino

Sensory evaluation has been defined as a scientific method used to evoke, measure, analyse, and interpret those responses to products as perceived through the sense of sight, smell, touch, taste, and hearing [8]. In many of these sensory evaluations the perception of consumers of food products is determined. Several sensory attributes of food products – for example creaminess, sweetness and bitterness – are scored by a panel of consumers. These attributes can be measured on various scales. The most frequently used scales are line-marking scales, category scales and Just About Right (JAR) scales. By a line-marking scale the consumer is asked to make a mark on a line to indicate the intensity of a sensory attribute. By a category scale the consumer choses between discrete response alternatives (for example a 9-point-scale: "not at all sweet" to "very sweet"). Just About Right scales are used to measure the *desirability* of a sensory attribute [9]. The JAR scale is bipolar measurement. In JAR scaling, two opposite anchors, for example "not sweet at all" and "much too sweet", are placed at each end of the scale, and the midpoint is labelled "just about right" [10].

A data set for the sensory evaluation of 5 different cappuccinos is used as a case study (Cappuccino case). In Figure 3 a part of the data set is given in Excel format. To summarize the experiment: each of 50 consumers taste all 5 different cappuccinos (products 195, 295, 374, 392 and 497), which they rate for 5 sensory attributes (appearance, overall liking, aroma, flavour and mouth feel) at a 9-point category scale. A 5-point Just About Right scale (-2 to 2) is used to measure the desirability of 4 sensory attributes (aroma, sweet, bitter and creamy).

consumer	product	appearance	overallliking	aroma	flavour	mouthfeel	aroma2	sweet	bitter	creamy
1	195	9	5	5	5	7	-2	0	0	0
1	295	7	7	6	8	7	0	-1	2	1
1	374	9	3	4	2	3	-2	0	0	-1
1	392	8	5	5	5	6	2	0	-1	-2
1	497	8	3	4	2	7	2	0	1	2
2	195	5	6	6	6	6	2	-2	1	2
2	295	4	6	6	6	6	0	2	0	2
2	374	6	3	3	3	3	0	0	2	0
2	392	6	6	6	6	6	0	-1	-1	-1
2	497	5	4	3	4	4	-1	-1	-1	0
3	195	6	7	5	7	6	-1	1	-1	-2

Figure 3 Part of sensory data set to evaluate 5 different cappuccinos

Our colleagues at the Consumer Science department of Food and Biobased Research make spider plots (also called radar charts) of the variables which are measured at JAR scales. They would like to visualize how the 5 cappuccino products deviate on average from the ideal product, i.e. the case in which all attributes score a "just about right" (score of zero).

2.1 Current workflow in Excel

The left part of Figure 4 shows the current workflow for making a spider plot in Excel. The complete data set is stored or imported in Excel. For this specific task, only the attributes at JAR scale need to be identified and selected for analysis. This can be relatively easy when the attributes at JAR scales are clustered in successive columns, as in the data set in Figure 3. However, often these attributes are scattered over the entire data set. Selecting and copying by hand is error prone and labour intensive. When the relevant JAR data is selected, the mean scores

for each attribute at JAR scale are calculated using Excel's functions and stored in a separate table with mean values. Thereafter, these means are visualized in Excel by a spider plot. This spider plot can be copied in a report or stored as a picture.

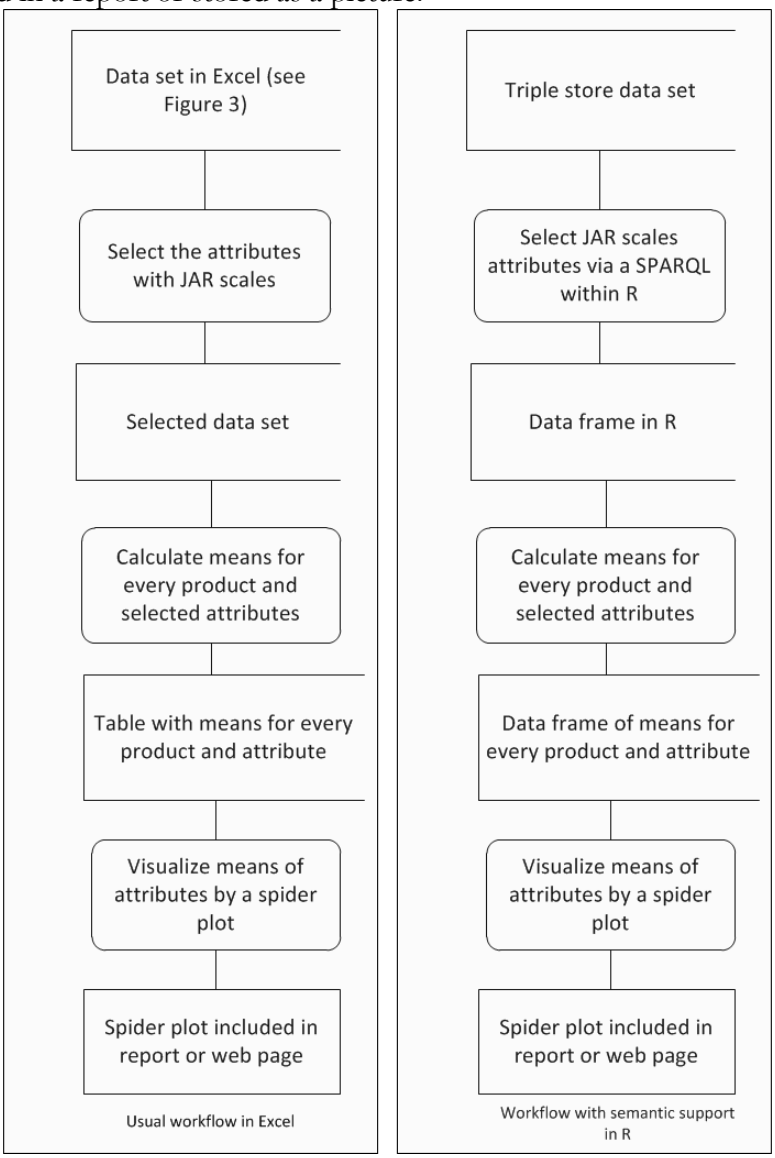


Figure 4 Workflow of visualizing attributes with JAR scales by a spider plot in Excel (left part) and with semantic support in R (right part).

2.2 Workflow with semantic support and R

The right hand side of Figure 4 shows the developed workflow with semantic support in R. The experimental design and the data are stored in a data model. The data model is defined in RDF/RDFS/OWL and contains several classes and properties. It reuses existing data models, such as OM and OQR.

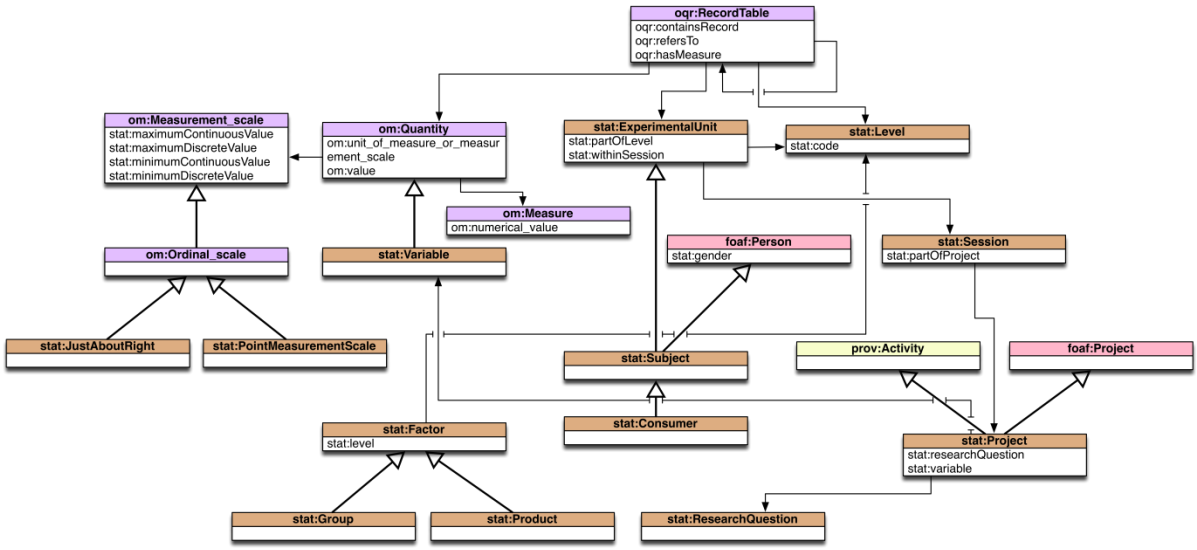


Figure 5 The Statistics Ontology and Data model.

The data model (see Figure 5) is defined in the statistics ontology, which contains classes to represent levels (e.g. products being tested), experimental units (e.g. subjects), variables, factors, measurement scales (e.g. Just About Right), projects, research questions and sessions. Where possible, it reuses concepts defined in other ontologies such as foaf:Person or foaf:Project in the Friend of a Friend (FOAF) ontology or prov:Activity from the Provenance Ontology (PROV). Most importantly, it reuses concepts in two ontologies that were developed at Food & Biobased Research for Units and Measures (OM) and quantitative research (OQR). OQR defines a set of concepts and properties that can be used to represent data in tabular and OM provides classes and properties used to represent for instance measurement scales such as ordinal scales (e.g. Just About Right, or Nine point scales).

Figure 6 presents a part of the RDF graph that results from importing an Excel sheet from the Cappuccino case. This figure only shows one row with two variables (sweet cappuccino flavour and cappuccino appearance), a subject (cappuccino consumer 3), and a product (cappuccino 295) is represented. This data representation seems very verbose, although it should be remembered that many of the nodes in this graph need only to be defined once and can be reused in multiple rows. The subject, for instance, needs only to be defined once for every subject. Another example are the values for the JAR scale (e.g. stat:Value0OnJustAboutRightFivePointScale with numerical value 0), these values can be reused in every table cell with a JAR scale with value 0.

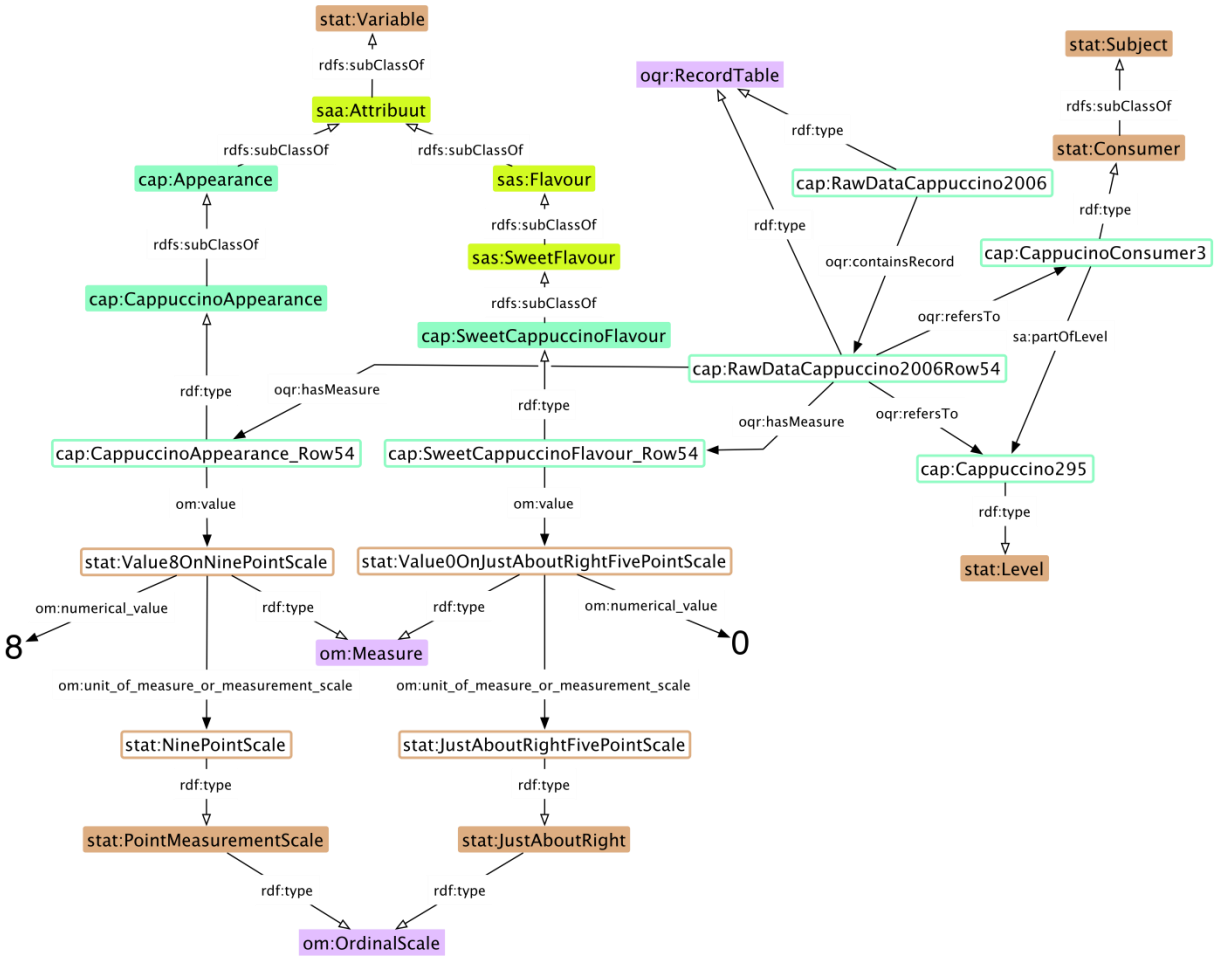


Figure 6 A graph representing the data from the Cappuccino case, for one specific row with two data values, a level (the cappuccino being tested) and a subject.

The data from the Cappuccino case were originally stored in an Excel file and transformed using a script into RDF data using the data model described above. The RDF data was then stored in a Sesame triple store, which provides a SPARQL endpoint for querying the data. The R SPARQL package [7] can query the triple store using a SPARQL query and transform the results into R data structures for further processing.

The triple store not only stores the attribute types and the values of the measurements (i.e. the data, as in Figure 3) but also information about the scales of the attributes being used, the attributes with JAR scales can be selected. In the data set there are two attributes measuring aroma, one at 9-point scale and one at JAR scale. The Excel table automatically changes the attribute "aroma" at JAR scale is named into "aroma2" because it occurs twice in the table header. Within a triple store the two "aroma" attributes are unique by the semantic annotation of these variables, each with their specific *unit_of_measure_or_measurement_scale* (see Figure 6).

The following SPARQL code is used to select the JAR scale attributes within the triple store.

```
# The prefixes used
```

```
PREFIX rdfs:<http://www.w3.org/2000/01/rdf-schema#>
PREFIX sa:<http://www.foodvoc.org/resource/statistical-
analysis/>
PREFIX cap:<http://www.foodvoc.org/statistics/cappuccino/>
PREFIX xsd:<http://www.w3.org/2001/XMLSchema#>
PREFIX owl:<http://www.w3.org/2002/07/owl#>
PREFIX om:<http://www.wurvoc.org/vocabularies/om-1.8/>
PREFIX rdf:<http://www.w3.org/1999/02/22-rdf-syntax-ns#>
PREFIX oqr:<http://www.wurvoc.org/vocabularies/oqr/>
PREFIX :<http://www.foodvoc.org/resource/statistical-analysis/>

# Subject here is the human participant of the experiment
# Product is the Cappuccino being tested
# VariableType is the variable (e.g. Sweet cappuccino flavor)
# Value is the measurement
# Scale is the type of Scale, in this case only JAR scales
SELECT DISTINCT ?subject ?product ?variableType ?value ?scale WHERE {
     # ?s is a row in the data table
     ?s a oqr:RecordTable.
     ?s oqr:refersTo ?subject.
     ?s oqr:refersTo ?product.
     # A subject needs to be of type :Subject
     ?subject rdf:type/rdfs:subClassOf* :Subject.
     # A product needs to be of type :Level
     ?product rdf:type/rdfs:subClassOf* :Level.
     ?s oqr:hasMeasure ?variable.
     # Each row has its own variable, we are interested in the
     # type of variable, e.g. Sweet cappuccino flavor.
     ?variable rdf:type/rdfs:subClassOf* :Variable.
     ?variable a ?variableType.
     ?variable om:value ?val.
     # ?val is not a numerical value, but an instance that also
     # contains for instance the scale.
     # Using the om:numerical_value property we get the numerical
     # value.
     ?val om:numerical_value ?value.
     ?val om:unit_of_measure_or_measurement_scale ?scale.
     # Here we select only variables that use a JAR scale.
     ?scale rdf:type sa:JustAboutRight.
}
```


In R, the data returned from the triple store is saved in a so-called data frame. This data frame can be used to apply several functions on these data. There are many functions in several packages that can be used to calculate the means of each product for each attribute over all consumers. For this study the package "psych" [11] is used to calculate summary statistics. The means are selected and stored in a data frame. Given this data frame a spider plot can be made. In this study the package "fmsb" [12] is used.
Summarizing, with a data model the experimental design and the measurements are stored in a triple store. This triple store can be approached by R using SPARQL. As soon as the data is imported in a data frame in R all kinds of computational operations can be performed at these data. When a researcher decides that a spider plot of JAR scale attributes is needed, selection of the attributes and generating the graph can be fully automated.

IV. Conclusion

In this study is a first step in investigating how semantic support can assist researchers with their data analysis. The data analysis should always be connected to the design of an experiment. This study shows that semantic support can provide the necessary connection between experimental design and data analysis.
Concepts of an experimental design are modelled in an ontology and stored in a data model. By guiding the researcher systematically (for example via a user interface) through the (statistical) choices of an experimental design the researcher can be prevented from using a wrong design. Furthermore, the choices made in the experimental design are integrated with the measurements of the experiment by semantic support. The data is not stored in a spreadsheet, but in a triple store where each data value is tagged with a unique identifier. For each data value the experimental design can be recovered providing for robust provenance information.
This case study demonstrates that with semantic support a selection of variables with a particular type of scale can be done easily. This is followed by an automation of further data analysis within R (or any other program). In the future we would like to extend this work with an interface to guide the researcher within sensory sciences in the statistical choices at the phase of the experimental design. These choices are stored in a data model as described in this study. Based on this data model a more comprehensive automatic data analysis of sensory evaluation of food will be developed.

3 References

[1] www.w3.org

[2] M. O'Mahony (1986). Sensory Evaluation of Food. Statistical methods and procedures. Marcel Dekker, New York.

[3] W3C (2014), RDF, Resource Description Framework, http://www.w3.org/RDF/

[4] Timothy Lebo, Satya Sahoo, and Deborah McGuinness (eds), W3C (2013), PROV-O: The PROV Ontology, http://www.w3.org/TR/prov-o/

[5] Eric Prud'hommeaux, and Andy Seaborne (eds), W3C (2008), SPARQL Query Language for RDF, http://www.w3.org/TR/rdf-sparql-query/

[6] W. N. Venables, D. M. Smith and the R Core Team (2013). An Introduction to R - Notes on R: A Programming Environment for Data Analysis and Graphics - Version 3.0.2.

[7] W.R. van Hage (2013). Package 'SPARQL', http://cran.r-project.org/.

[8] H. Stone and J.L. Sidel (1993). Sensory Evaluation Practices, 2d ed. Academic, San Diego.

[9] H.T. Lawless and H. Heymann (1998). Sensory Evaluation of food – Principles and Practices, Chapman & Hall, New York.

[10] L. Rothman and M.J. Parker (2009). Just-About-Right (JAR) scales: Design, usage, benefits, and risks. American Society for Testing & Materials.

[11] W. Revelle (2014). Package 'psych', http://cran.r-project.org/.

[12] M. Nakazawa (2014). Package 'fmsb', http://cran.r-project.org/.

[13] 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), 201, pp.276–287.

