Identifying and Extracting Quantitative Data in Annotated Text*

Don J.M. Willems[1], Hajo Rijgersberg[1], and Jan L. Top[1,2]

[1] Wageningen UR, Food and Biobased Research
[2] Dept. of Computer Sciences, Vrije Universiteit Amsterdam

Abstract. In science it is difficult to reuse quantitative scientific data. For example, it is not possible to search for quantitative data in papers in a directed way, such as using the query "Select the storage modulus of dairy product A after the temperature has decreased from 90 to 4 °C". This is caused by the fact that data is made available in (relatively) free formats as in scientific papers, spreadsheets, or databases, all with limited annotation and description of the way they were obtained. Meaning is lost, for example about what the numbers relate to (quantities and units are often poorly indicated). Many researchers, especially in the physical and computer sciences use LaTeX in their creation of scientific papers. In this paper we present a set of LaTeX-style files, which use the terminology defined in wurvoc.org, that can be used to annotate scientific papers. These style files define a set of commands, each representing a specific quantity or unit. If the LaTeX is typeset into a PDF file, quantities and units in the PDF will be annotated with the appropriate references (URIs) to the corresponding concepts in the OM ontology. This will not only disambiguate the use of these quantities and units, but will also enable us to extract triples from the PDF, facilitating the use of SPARQL queries to answer advanced quantitative question.

1 Introduction

Many scientific papers are written using the LaTeX typesetting system and published as PDF. It is desirable to process this knowledge automatically. We propose a method to add semantic annotations to LaTeX files, extending the typesetting methods that LaTeX offers. Using these annotations we can automatically extract information from the generated PDF files.

In scientific research one is often looking for quantitative data. One can find these for example in external databases and spreadsheets or in scientific papers. However, it is difficult to look for quantitative data on the Web in a directed way. For example, the query "Select the storage modulus of dairy product A after the temperature has decreased from 90 to 4 °C" can not be carried out because computer tools are not able to link the correct numbers to the correct units and correct quantities. Another example is to give maximum, minimum and average values observed for parameter X in a set of papers. The problem is caused by the fact that data are expressed in

* This publication was supported by the Dutch national program COMMIT.

relatively free formats, such as in text or in spreadsheets. The structure of the data is often not clear for a computer. A lot of research is being done on parsing the structure of scientific papers including tables (see for instance [1]), which often contain a lot of quantitative data. The annotation of the data, however, is usually limited, including quantities and units that are often poorly indicated [2]. In other words, the context to which the numbers refer is often lost.

To approach these problems, formal terminology is developed in the discipline of the Semantic Web [3]. This terminology can be made publicly available through the Web, so that it can be used (referred to) in digital sources [4]. The ideal situation would be to have all information available in formal languages. This, however, requires a major effort in creating a consistent conceptual model of scientific knowledge and significant advances in automatic parsing and annotation of scientific texts. The work presented in this paper represents an important step in annotating texts with formal concepts.

Often quantitative work, in for instance the physical sciences, is presented in scientific papers that were created using the LaTeX typesetting system[5]. LaTeX is especially suited for the physical and computer sciences because of the extensive support for mathematical typesetting. The use of LaTeXstyle files provides a convenient way to add annotations to content, not only to the main text of an article but also to tables and even graphics created in LaTeX.

In this exercise we focus on the annotation and extraction of quantities and units, defined in our Ontology of units of Measure and related concepts (OM; see Section 3) from annotated content. Using custom LaTeX commands, quantities and units with semantic annotations are inserted into the PDF with the appropriate references (URIs) to the corresponding concepts in the OM ontology. This will not only disambiguate the use of these quantities and units, but will also enable us to extract RDF [6] triples from the PDF, enabling the use of SPARQL [7] queries to answer advanced quantitative questions. This may greatly enhance searching and the processing of data in general. Because a computer will suffer less from ambiguity owing to the annotations, it can reach a higher quality in the support that it offers.

In this paper we will first focus on related work (Section 2). Subsequently, in Section 3, we will briefly describe OM, discuss aliases in LaTeX, propose commands for referring to quantities and units in LaTeX, and describe the method for transforming equations from LaTeX to RDF. In Section 4 we will present a few examples of annotated scientific texts and in Section 5 we will discuss the results.

2 Related work

General developments in the area of semantic publishing, have been presented in several papers [8,9]. The authors express the need for annotating publications and extracting structured information from them. Otherwise the sheer number of references hinders interaction between related scientific activities. Assessing and integrating previous work benefits significantly if factual information is (also) available as Linked Open Data. These authors indicate the need for explicating the structure of arguments in scientific discourse, as well as a structured presentation of the con-

cepts and relations between concepts. Our work is complementary, in the sense that we start by disclosing numerical facts which can be related to other concepts. We take a pragmatic approach by building on standard practice in writing LaTeX documents.

Some approaches exist for semantically annotating LaTeX files. STEX, Semantic Markup for TeX/LaTeX [10], consists of a collection of TeX macro packages that allows the user to markup TeX/LaTeX documents semantically, turning the documents in a format for mathematical knowledge management (MKM). The method focuses on mathematical relations, collections, and formats of numbers (e.g., decimals). STEX, however, cannot be used to annotate quantities and units.

The SIunits package for LaTeX [11] is a package that provides support for typesetting units, in more or less the same way as we do. Quantities, however, do not appear as such in this package. Moreover, as the package only provides typesetting, the concepts are not linked to a centrally available vocabulary on the Semantic Web.

SALT, Semantically Annotated LaTeX[12], provides a means for externalising rhetorical and argumentation captured within a publication's content. However, the approach does not relate to mathematical concepts or quantities and units.

Mathematics can be expressed in XML using MathML [13] and can as such be incorporated into web pages. Equations in MathML can be expressed in two distinct formats: i) Presentation encoding, which as the name suggest supports the construction of traditional mathematical notation. ii) Content encoding supports the "encoding of the underlying mathematical meaning of an expression" [13]. While the scope of MathML itself does not include units, they can, however, be expressed in MathML [14], including a reference to the URI of a concept in a formal vocabulary of units (such as OM).

Ideally, structured information should be extracted automatically from publications. Frameworks such as GATE [15] provide functionality for automatically annotating text. It does not, however, provide the ability to automatically annotate equations, where only symbols are used for quantities and units, or graphics.

In previous work we have annotated quantities and units in Excel files, using an add-in for Excel we developed along with web services disclosing quantities and units from OM, and operations that can be performed on these terms, such as dimension and unit consistency checks on formulas and returning possible units for a particular quantity. In the present work we relate terms in LaTeX documents to this same ontology (OM [16]), extending the use of OM concepts to a broader audience.

3 Method

Typesetting (the creation of a visual representation of a text) using LaTeX is done using the TeX typesetting engine [17] developed by Donald Knuth in the 1970s and 1980s. TeX provides a set of low level declarations or *commands* for typesetting. LaTeX, developed by Leslie Lamport in the 1980s, provides a set of higher level commands that can be used to easily create documents without having to worry about their typographical appearance [5]. These sets of commands can easily be extended (and often are) by users to define a set of personal commands for typesetting particular pieces

of information that are often used. These commands are either defined in the main LaTeX source file or in style files which can be imported into the main LaTeX file.

In this paper we present a LaTeX package (as a set of style files) that uses terminology as offered through our ontology platform `wurvoc.org`.

3.1 Ontology of units of Measure and related concepts (OM)

The ontology that we use to refer to in the LaTeX files is OM. OM is an ontology based on older ontologies of units of measure, such as EngMath by Tom Gruber [18]. In earlier work [16] we have compared OM, EngMath and other ontologies, and OM appeared to be the most extended ontology, e.g. defining the most of the relevant concepts in the quantitative domain, such as "quantity", "unit of measure", "dimension", "measure", "measurement scale", etc.

OM defines concepts such as unit, quantity and dimension. Quantities are related to units of measure and measurement scales that can be used to express them using the relation `\unit_of_measure`. Units of measure are defined by some observable standard phenomenon, such as the length of the path travelled by light in a vacuum during a time interval of 1/299 792 458 of a second, for meter. Measures, such as "3 kilogram" are used to indicate values of quantities. Multiples and submultiples of units have a prefix, such as in kilogram and millimetre.

Systems of units organise quantities and units of measure, e.g. the International System of Units (SI). Such a system defines base units and derived units. Base units are units that cannot be defined in terms of other units (e.g. metre and second). Base units can be combined into derived units, such as for example metre per second ($\mathrm{m\,s^{-1}}$).

OM is based on a semi-formal description of the domain of units of measure, drafted from several paper standards that we have analysed, e.g. the Guide for the Use of the International System of Units [19], by the NIST. For a full list of statements, the sources that we have used, and ontological choices made, see previous work [16].

OM is modelled in OWL 2 [20]. The ontology is published as Linked Open Data [21] through our vocabulary and ontology portal wurvoc.[3] OM can be used freely under the Creative Commons 3.0 Netherlands license.

3.2 Aliases in LaTeX

When using LaTeX it is often preferable to create *aliases* for often used (complex) command structures instead of retyping these command structures again and again.

For instance, LaTeX source code becomes more difficult to interpret when units are used in an equation. To create a statement like:

$$G = 6.673 \times 10^{-11}\,\mathrm{N\,m^2\,kg^{-2}} \tag{1}$$

which is the gravitational constant, the following LaTeX source code can be used:

[3] http://www.wurvoc.org/vocabularies/om-1.8/. The objective of wurvoc.org is to publish vocabularies and associated web services relevant to the general domain of physical units and quantities and in particular the domains of life sciences and agrotechnology. In wurvoc one can browse vocabularies and directly interface with them.

```
G = 6.673\times 10^{-11} \mathrm{N} \mathrm{m^2} \mathrm{kg^{-2}}
```

Units are written in a non-bold but upright font (i.e. not cursive as is used for quantities and variables).

To make things easier, authors using LaTeX construct self-defined aliases. For instance, for the unit for the gravitational constant we might define a new command:

```
\newcommand{\Gunit}{\mathrm{N} \mathrm{m^2} \mathrm{kg^{-2}}}
```

and for exponents:

```
\newcommand{\E}[1]{\times10^{#1}}
```

The author can then use his custom defined commands (or aliases) `\Gunit` and `\E` to insert the correct unit and exponent. Equation 1 can then be typeset using

```
G = 6.673\E{-11} \Gunit
```

which is much easier to interpret by humans.

Sets of often used aliases can be created and distributed using style files. These LaTeX examples use custom commands to provide easier typesetting of mathematical expressions. We would like to use these typesetting commands (aliases) to insert semantic information into the mathematical expressions.

3.3 Semantic annotations

As aliases are used quite often by authors, it becomes possible to add extra information to the output produced when typesetting LaTeX files. The extra information we would like to provide in scientific texts are links (URIs) to ontological definitions of the quantities and units used as defined in the Ontology of units of Measure (OM, [22], prefix is `om:`). To this end we have created a set of style files (a package) that define a large set of aliases that not only create the correct symbols and layout for quantities and units, but also provides annotated links to the ontological concepts describing these quantities and units. As most LaTeX source files are typeset into PDF files these days, we have decided to use PDF annotations (more specifically hyperlinks) to create the links to the ontological concepts. The URI of the annotated concept is added to the PDF as a hyperlink and will generally only be visible when the mouse cursor hovers above the linked text.

Using the hyperref package, hyperlinks are inserted into the PDF produced by LaTeX by defining aliases with a custom command:

```
1  \newcommand{\annot}[3]{
2     \ifthenelse{\isempty{#3}}
3       {\href{om:#1}{#2}}
4       {\href{om:#1}{#3}}
5  }
```

In the first line, the command `\annot` is defined with three parameters. The first parameter is the part of the URI in the OM ontology of the concept (quantity or unit) that comes after `om:`, which is the base URI for the OM ontology. The second parameter is the default symbol used for this quantity or unit, and the third (optional) parameter can be used by an author to insert a custom symbol for the same concept. The second line checks whether the author has used an optional symbol. If not, the third line will insert the default symbol (the second parameter) with a hyperlink consisting of the base URI concatenated with the first parameter. If the author used the third parameter for a custom symbol, this symbol is used instead, with the same URI (line 4).

The `\annot` command is defined in the `om.sty` style file provided in our OM-LaTeX distribution. A few other commonly used commands, such as `\vect` to typeset vectors, `\E` to typeset exponents such as 4.2×10^3, and `\unit` to typeset units are also provided in this file.

To annotate an equation like:

$$||\mathbf{a}|| = 5.433 \times 10^{-1}\,\mathrm{m\,s^{-2}} \tag{2}$$

which would normally be produced by the source code:

```
||\vect{a}|| = 5.433 \times10^{-1}\unit{m}\unit{s^{-2}}
```

can now be obtained, with the same result, using the following code:

```
||\Acceleration || = 5.433\E{-1}\metrePerSecondSquared
```

This LaTeX code, while not much shorter, is more easy to interpret by humans. For all units and quantities in OM a human readable alias, such as `\Acceleration`, is provided in the LaTeX style files. Aliases from the SIUnits package [11] will also be included, so that texts created with SIUnits can easily be converted to include OM annotations.

More importantly, however, for our purposes, is the addition of the hyperlink pointing to the relevant concept. To facilitate this, the following aliases were defined:

```
1  \newcommand{\Acceleration}[1][]{
2     \annot{Acceleration}{\vect{a}}{#1}
3  }
4
5  \newcommand{\metrePerSecondSquared}[1][]{
6    \annot{metre_per_second_squared}{\unit{m} \unit{s^{-2}}}{#1}
7  }
```

In line 2, we use the `\annot` command to create an alias for the quantity acceleration with URI: `om:Acceleration` and default symbol '**a**'. In line 6 the same is done for the unit metre per second squared (URI: `om:metre_per_second_squared`). The first parameter to the `\annot` command (`Acceleration` in line 2, and `metre_per_second_squared` in line 6) provide semantic annotations to the mathematical expression. The second (`\vect{a}` in line 2, and `\unit{m} \unit{s^{-2}}` in line 6) and third parameters (`#1` in both line 2 and 6) are only concerned with typesetting.

All LaTeX commands representing quantities and units can also be used with user defined symbols simply by adding an (optional) parameter to a command. For instance, the command `\LuminousFlux` produces the symbol for the quantity luminous flux 'F' with a link to the related concept (`om:Lumious_flux`) in the OM ontology. If the author wants to use another symbol to represent luminous flux, he or she can achieve this by specifying the alternative symbol as an argument: `\LuminousFlux[\Phi]` produces 'Φ', still linked with the same concept in the OM ontology. If desired sub- and superscripts can also be used in the argument: `\LuminousFlux[F_{\lambda}]` produces 'F_λ', again linked with the same URI.

3.4 URI and equation extraction

When using the typesetting tool `pdflatex` to create PDF files from the LaTeX source, the URIs representing the unit and quantity concepts are inserted as hyperlinks into the PDF. To use these annotations we have to parse the PDF files to find the hyperlinks (URIs). Using Apache's PDFBox `http://pdfbox.apache.org/` we were able to create a small Java tool to parse the PDF files and extract the URIs representing concepts in OM and linking these URIs to the text.

Using this setup we are able to extract OM concepts (units and quantities) from a text generated with OM-annotated LaTeX. We would, however, also like to extract the semantics of statements like $V = 15.2\,\mathrm{m}^3$ (i.e. we would like to extract the fact that the quantity *volume* has a value of 15.2 in units of *cubic metre*). To this end we have also added the functionality of finding binary (=, <, >, ≈, etc.) relations in the text to the extraction tool.

When a PDF is parsed by the extraction tool, URIs for units and quantities, numeric values and binary relations are tagged in the text. Operators, like `\E` are also recognised, and in the case of exponents, the value is changed accordingly (e.g. 5.2×10^3 is changed to 5200). The tool then applies rules to find patterns in the text like:

```
[QUANTITY] [BINARY_RELATION] [VALUE] [UNIT]
```

If the tool comes across such a pattern, the combination of quantity, relation, value, and unit is stored. For instance the equation:

$$E_\mathrm{k} = 1.209 \times 10^{-2}\,\mathrm{eV} \tag{3}$$

is extracted as:

```
1 [QUANTITY=om:Kinetic_energy] [BINARY_RELATION='=']
2        [VALUE=0.01209] [UNIT=om:electronvolt]
```

In this manner quantitative statements can be extracted from PDF generated with OM annotated LaTeX.

3.5 Transformation to RDF

The result of the extraction can then be transformed into RDF statements using the OM ontology. For instance, the following equation

$$F = 15.2\,\mathrm{N} \tag{4}$$

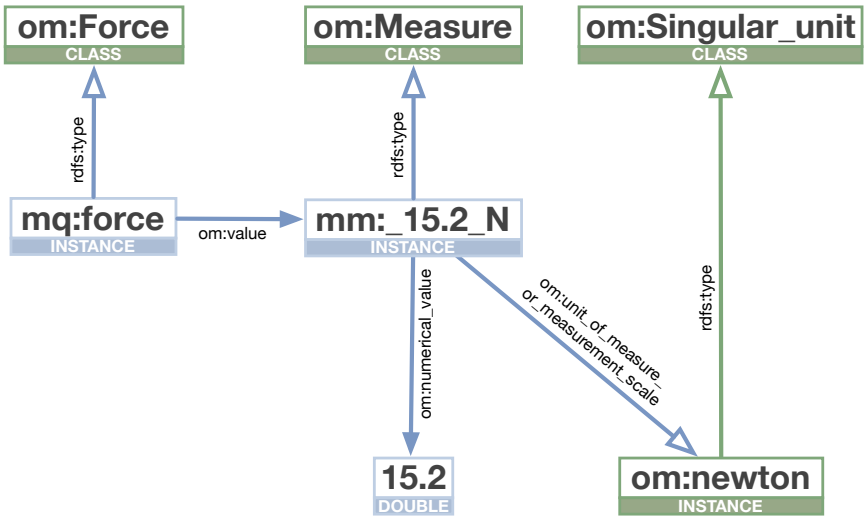

Fig. 1. Extracted RDF representing Equation 4. This graph only represents OM-specific data; other information such as provenance data are present in the full RDF graph.

can be transformed into RDF (in turtle format [23]):

```
mq:force om:value mm:_15.2_N;
    a om:Force .
mm:_15.2_N a om:Measure ;
    om:numerical_value "15.2"^^xsd:double ;
    om:unit_of_measure_or_measurement_scale om:newton .
```

where `mm` and `mq` are prefixes for custom defined namespaces (possibly pointing to the URI for the original text, thereby ensuring provenance) for measures and quantities respectively, and `om` is the prefix for the OM namespace. This statement can also be visualised as a graph (Figure 1).

The current extraction tool is not only able to create the statements to model the equation in RDF, but it is also able to export these RDF statements to an RDF triple store, where it can be combined with other semantic data extracted from the PDF, or obtained from other sources.

4 Real-world examples

The number of detected measures depends on the type of paper; experimental papers tend to contain more measures than theoretical papers. For example the fifth page of a paper on water vapour sorption in gluten and starch films contains the following text:

> [...] The obtained parameter values for starch films ($T_g = 540 \pm 10\,\mathrm{K}$ and $\Delta C_p = 0.32 \pm 0.02\,\mathrm{J\,g^{-1}\,K^{-1}}$) differ from the values obtained by van der Sman and

> Meinders (2011) from the data of starches from different botanical sources and obtained with different experimental techniques ($T_g \approx 475\,K$, $\Delta C_p \approx 0.43\,J\,g^{-1}\,K^{-1}$) but are in close agreement with the values obtained by Bizot et al. (1997) for pea amylose determined using DSC ($T_g = 589\,K$ and $\Delta C_p = 0.27\,J\,g^{-1}\,K^{-1}$). [...][4]

which contained six measures. Our extraction tool extracted all six measures correctly, for example, the following RDF triples were extracted for the last measure of change of water specific heat ($\Delta C_p = 0.27\,J\,g^{-1}\,K^{-1}$):

```
1    mq:specific_heat om:value mm:_0.27_JpgK;
2        a om:Specific_heat.
3    mm:_0.27_JpgK rdfs:type om:Measure;
4        om:numerical_value "0.27"^^xsd:double;
5        om:unit_of_measure_or_measurement_scale
5                om:joule_per_gram_kelvin.
```

As one can see the measure has been extracted successfully and is correctly modelled in terms of OM. At the moment, we assume that ΔC_p is one symbol for a specific quantity and not a mathematical operation. Finally, it is not possible to add error values to the conceptual model (e.g. $T_g = 540 \pm 10\,K$) and to distinguish between = and $\approx$.

The extracted triples do not specify the source of the specific heat (i.e. the specific heat for pea amylose determined using DSC). This is a case for further (automatic) semantic annotation beyond OM. Including more extensive annotation would make the semantic information even more valuable. One could start searching in scientific RDF databases for articles on "specific water vapour heat" with a value between "0.2" and "$0.3\,J\,g^{-1}\,K^{-1}$".

As a second example of measures in an experimental paper consider the abstract of a paper on observations of a young star. The abstract alone contains six measures:

> We present CS(J=2-1) interferometric observations obtained with the Nobeyama Millimeter Array (NMA) toward a protostar (GH2O 092.67 + 03.07) in the Cygnus OB7 giant molecular cloud (distance = 800 pc). The data clearly indicate the presence of a compact (size $\simeq 8 \times 10^3$ AU) and young out-flow with dynamical time scale $\simeq 3500$ year. [...] We derive a total mass of $\simeq 0.6\,M_\odot$ and $\simeq 12\,M_\odot$ for the outflow and disk respectively. The comparison of the NMA data with a simple model of infalling disk indicates a mass of the central object in the range $4.0 < M < 7.5\,M_\odot$. [...][5]

In this example the names of the quantities are annotated as text (not math, e.g. 'size $\simeq 8 \times 10^3$ AU') as in: `\Diameter[size]` and is actually parsed correctly by our extraction tool:

[4] Laura Oliver, Marcel B.J. Meinders, Dynamic water vapour sorption in gluten and starch films, Journal of Cereal Science, **54**-3 (2011), pages 409-416.

[5] Bernard, J.P., Dobashi, K. , Momose, M.: Out flow and disk around the very young massive star GH2O 092.67+03.07. Astronomy and Astrophysics **350** (1999), pages 197-203.

```
mq:size om:value mm:_8000_AU;
    a om:Diameter.
mm:_8000_AU rdfs:type om:Measure;
    om:numerical_value "8000.0"^^xsd:double;
    om:unit_of_measure_or_measurement_scale
              om:astronomical_unit.
```

Please note that the numerical value containing an exponent (8×10^3) is interpreted correctly (8000).

5 Discussion

Embedding numerical facts in otherwise textual documents incurs a tension between the use of natural language and structured formats. We submit that scientists should be able to put their arguments forward with minimal technical constraints. On the other hand, embedding RDF-OWL type annotations eliminates ambiguity and simplifies computer processing. For example, consider the following statement:

> The water vapor permeability for optimal crispness and crumb softness retention was 8×10^{-9} g/(m s Pa).[6]

It is possible to request the author to annotate individual quantities and units of measure (and concepts), but it would also be possible to have the author provide RDF triples for the entire sentence. The first option seems less attractive from a computer processing perspective. In that case, more effort is required to parse the information into an equivalent RDF triple afterwards. Nevertheless we choose to stay close to normal writing as much as possible. By annotating at the level of quantities and units only, precisely enough formalisation is provided to enable automated construction of the composite triple. Moreover, by using the alias mechanism provided by LaTeX and our definition of `\annot`, the natural language style is approximated as much as possible.

This paper describes how units and quantities can be annotated. However, the value of such annotation is limited if it is not clear to which objects or phenomena these quantities refer. For example, $V = 15.2\,\mathrm{m}^3$ only becomes a useful fact if we know that it refers to a container containing water, or even to a specific container in an experiment. This would require annotating objects and phenomena using domain-specific ontologies, and relating them to the quantities used. A simple generalisation of our approach is to include the full URI in the `\annot` construct. This allows the user to link any object to an ontological class or instance. However, some heuristic processing would still be needed to link these objects to the annotated quantities. We consider this a necessary step in our method, but beyond the scope of the present paper.

[6] Anita Hirte, Rob J. Hamer, Marcel B.J. Meinders, Kevin van de Hoek, Cristina Primo-Martín. Control of crust permeability and crispness retention in crispy breads. Food Research International, Volume 46, Issue 1, April 2012, Pages 92-98

Finally we note that OM, the ontology of quantities and units, is accessible through a set of web services that provide additional functionality if data is annotated along the above lines. They allow automatic checking of combinations of units and quantities for correctness and completeness, but also automatic unit conversions. These can be useful aids during paper writing or reading.

6 Conclusion

For the research presented in this paper we have created a set of style files for LaTeX that refer to concepts from an ontology of units and quantities. By using the commands used in these style files, quantities and units are annotated directly. The concept's URI is included as a hyperlink when generating the PDF. Using these annotations we have been able to extract triples from the PDF and insert them into an RDF triple store, which can be queried with specific querying constraints. The LaTeX style files and corresponding PDF extraction tool will be made available in the near future.

In a broader sense it becomes feasible to do more with the annotated data, such as unit conversion, checking of dimension and unit consistency, integrating, performing computational methods on the data, etc. This functionality is available via OM web services [16]. To make the data even more reusable, it will be important to extract other concepts than quantities and units, such as the object or event that a value of such quantity refers to.

Ideally we should be able to annotate existing papers automatically. Frameworks such as GATE [15], which provide automatic annotation will play an important role in this endeavour. In earlier work [2] we have drafted heuristic rules for interpreting and formalising quantitative information in spreadsheets. This research could be extended towards quantitative information in scientific papers. At this moment, however, automatic annotation of measurements cannot be performed reliably enough in the cases we observed, which are intrinsically ambiguous and incomplete [2]. So, manual (and therefore, user-validated) annotation by authors is still required. The described method in this paper helps to annotate quantitative concepts such as quantities and units of measure, using the embedded URLs.

As the user will likely be using alias commands in LaTeX anyway, extending these with semantic annotations does not require extra effort for the user and these annotations are, therefore, relatively for free. The user only needs to include the OM-LaTeX package and is then able to use aliases with names close to the actual names of the units or measures, making the text easier to read.

In the light of extending this approach, we aim to investigate whether integration with STEX, SIunits, or SALT is possible. And as it is useful to annotate mathematical relations and operators, we will, moreover, define the URIs for relations and operators and more in OQR (Ontology of Quantitative Research) [24].

References

1. Oro, E., Ruffolo, M.: PDF-TREX: An Approach for Recognizing and Extracting Tables from PDF Documents. In: Proceedings of the 10th International Conference on Document Analysis and Recognition. (july 2009) 906 –910

2. van Assem, M., Rijgersberg, H., Wigham, M., Top, J.: Converting and Annotating Quantitative Data Tables. In Patel-Schneider, P.F., Pan, Y., Hitzler, P., Mika, P., Zhang, L., Pan, J., Horrocks, I., Glimm, B., eds.: Proc. 9th Int'l Semantic Web Conf. (ISWC'10). Number 6496 in LNCS, Springer-Verlag (2010)
3. Berners-Lee, T., Hendler, J.: Publishing on the Semantic Web. Nature (April 26) (2001) 1023–1025
4. Hey, T., Trefethen, A.: Cyberinfrastructure for e-science. Science **308** (2005) 817 – 821
5. Mittelbach, F., Goossens, M., Carlisle, J.B.D., Rowley, C.: The LaTeXCompanion, 2nd edition (TTCT series). Addison-Wesley, Reading, Massachusetts
6. W3C: Resource Description Framework (RDF). `http://www.w3.org/RDF/` (2004)
7. Prud'hommeaux, E., Seaborne, A.: SPARQL Query Language for RDF. `http://www.w3.org/TR/rdf-sparql-query/` (2008)
8. De Waard, A.: From proteins to fairytales: Directions in semantic publishing. IEEE Intelligent Systems (March/April) (2010)
9. Shum, S.B., Clark, T., de Waard, A., Groza, T., Handschuh, S., Sandor, A.: Scientific discourse on the semantic web : A survey of models and enabling technologies. Semantic Web Journal Interoperability Usability Applicability (Special Issue on Survey Articles) (2010)
10. Kohlhase, M.: Semantic Markup for TEX/LATEX. (2004)
11. Heldoorn, M.: The SIunits package. Consistent applications of SI units. (2007)
12. Groza, T., Handschuh, S., MŽller, K., Decker, S.: Salt - semantically annotated latex for scientific publications. Lecture Notes in Computer Science **4519** (2007) 518–532
13. Ausbrooks, R., Buswell, S., Carlisle, D., Chavchanidze, G., Dalmas, S., Devitt, S., Diaz, A., Dooley, S., Hunter, R., Ion, P., Kohlhase, M., Lazrek, A., Libbrecht, P., Miller, B., Miner, R., Rowley, C., Sargent, M., Smith, B., Soiffer, N., Sutor, R., Watt, S.: Mathematical Markup Language (MathML) Version 3.0. `http://www.w3.org/TR/MathML3/` (2010)
14. Harder, D.W., Devitt, S.: Units in MathML. `http://www.w3.org/TR/mathml-units/` (2003)
15. Cunningham, H., Maynard, D., Bontcheva, K., Tablan, V.: GATE: A framework and graphical development environment for robust NLP tools and applications. In: Proceedings of the 40th Anniversary Meeting of the Association for Computational Linguistics. (2002)
16. Rijgersberg, H., Wigham, M., Top, J.L.: How semantics can improve engineering processes: A case of units of measure and quantities. Advanced Engineering Informatics **25**(2) (2011) 276–287
17. Knuth, D.E.: The TeXbook. Addison-Wesley (1986)
18. Gruber, T., Olsen, G.: An Ontology for Engineering Mathematics. In: Fourth International Conference on Principles of Knowledge Representation and Reasoning, Morgan Kaufmann (1994)
19. Taylor, B.N.: Guide for the use of the International System of Units (SI). 2008 edn. Technical report, National Institute of Standards and Technology (2008)
20. W3C: Owl 2 web ontology language. Technical report, World Wide Web Consortium (W3C) (2009)
21. Bizer, C., Heath, T., Berners-Lee, T.: Linked Data - The Story So Far. International Journal on Semantic Web and Information Systems **5** (2009) 1–22
22. Rijgersberg, H., van Assem, M., Wigham, M., Broekstra, J., Top, J.: Ontology of units of measure (OM). `http://www.wurvoc.org/vocabularies/om-1.8/` (2010)
23. Beckett, D., Berners-Lee, T.: Turtle - Terse RDF Triple Language. `http://www.w3.org/TeamSubmission/turtle/` (2011)
24. Rijgersberg, H., Top, J.L., Meinders, M.: Semantic Support for Quantitative Research Processes. Intelligent Systems **24**(1) (2011) 37–46