Identifying Units and Quantities in Technical Documents

Mihai Cîrlănaru

Computer Science
Jacobs University Bremen
Campus Ring 1
28759 Bremen
Germany

Type: *Guided Research Proposal*
Date: *February 24, 2011*
Supervisor: *Prof. Dr. Michael Kohlhase*

Abstract

Precise and semantically accurate description of measurable concepts is extremely important when it comes to critical systems like medical diagnosis, airplanes and very expensive machinery like satellites. Units and quantities encompass the necessary apparatus for describing precision and, at a higher level of perception, provide a sense of magnitude. However, the wide range of different standards for representing units increases the chances of misinterpreting them. Technical documents provide a starting point for reasoning about units and quantities because of the high density of occurrences they include. The need of a consistent way of understanding and interacting with units and quantities is motivated by the diversity of their representations both syntactically and semantically. The intended research will focus on the occurrence of units and quantities in technical documents from the ARXMLIV corpus. The goal of the project is the successful identification and disambiguation of units and quantities represented in content markup languages for mathematical formulae like OPENMATH and *Content* MATHML. This will enable a great extent of new technologies and services that will provide methods for preventing hazardous behavior of critical systems.

1 Introduction and Motivation

Units and quantities are present *everywhere*.
Formally, a **unit** is *any determinate quantity, dimension, or magnitude adopted as a basis or standard of measurement for other quantities of the same kind and in terms of which their magnitude is calculated or expressed* [Oxf], but from the top-most level of perception, it simply provides information on a wide range of quantifiable aspects. This leads to a very high occurrence rate in one's daily life. For example, most of the general-use, common devices, like watches, phones, cars and so on, provide information (feedback) through the use of units and quantities. Descriptions, specifications, routines also take advantage of the ability to quantify information and also to provide a certain perception level through units. Concrete examples comprise of things as unrelated as cooking receipts, medical prescriptions, scientific papers and many other, thus emphasizing the great extent units and quantities possess. Having a (automated) way of identifying and understanding these occurrences can enable the evolution of impressive technologies and services (see section 4.5).

Units and quantities are a very *important* aspect of our daily lives.
At the perception level, aside from quantifying properties and relations between objects, units bring the meaning of scale. One of their most important uses was and still is in the field of science: they have allowed scientists throughout the past couple of centuries to better transmit and exchange knowledge between them.

Units and quantities come in very different formats and types. Be it just for the sake of simplicity (easier calculations, easier perception of their scales) or just for the sake of legacy (used for centuries), these types, although standardized, pose a lot of problems in real life. Consider losing a $125 million satellite [Mar] because of the differences between metric and imperial unit systems or running out of fuel in mid-flight with an aircraft whose fuel sensors were faulty configured in displaying the units [Air]; and these are just a few of the most notable errors caused by units. Fields like medicine, commerce, civil engineering have also been marked by such types of errors and pitfalls [Usm].

The high density of units and quantities in technical documents enables an extensive analysis of different context occurrences. This confers a broader view on their hidden semantics and also helps to understand their different aspects thus avoiding the common pitfalls and hazardous interpretations mentioned above [Usm]. Having a better understanding of the different types of occurrences of units and quantities would enable the development of a plethora of useful tools and services (for more details see section 4.5) that can considerably improve the ways of interacting with their semantics.

2 State of the Art

The main technologies in focus for the research project are the content markup languages OPENMATH and *Content* MATHML in the context of the ARXMLIV corpus (cf. section 2.2). Also previous research on units and related services is described in section 2.3.

2.1 OPENMATH and *Content* MATHML

In the past couple of years the web was enriched with an important number of XML-based, content-oriented markup languages for mathematics and natural sciences. The focus of the investigations discussed by this proposal will be on OPENMATH [Bus+04] and *Content* MATHML [Aus+10] (with a higher preference for the latter), which are standards for the representation and communication of mathematical objects [KR09]. Both markup languages are concerned with the encoding of object's semantics rather than with its visual representation, to enhance the ease of knowledge exchange and processing by software systems and also human beings. The higher interest in *Content* MATHML is justified by its preponderance in the pool of technical documents that will be analyzed (more details in section 2.2). Nevertheless there are means of converting one into the other [Car+01] which enables a great flexibility of selecting the most suited one for a respective type of technical document and also switching from one to another.

OPENMATH provides an extensible framework with well-defined extension mechanisms through Content Dictionaries (CDs) [Bus+04]. Their main purpose is to provide semantics for the symbols used by OPENMATH objects. This modular aspect allows an easy expansion of the unit knowledge base [SD08] to further disambiguate units in all the possible common occurrences they might have. OPENMATH's minimal core language enables a high degree of portability and extensibility which can only be an advantage for the study of units and quantities (more details in section 3.1).

In a similar way, *Content* MATHML is concerned with the representation of object semantics through an extensive collection of construction elements whose intuitive format covers all of school and engineering mathematics (the "K-14" fragment) [KR09]. *Presentation* MATHML is the other sublanguage of MATHML and is concerned with the layout schemata for describing the two-dimensional notation of mathematical formulae and will not be treated in this proposal (for more details see section *Presentation Markup* of [Aus+10]).

Structure-wise, both OPENMATH and MATHML provide a valuable basis for machine processing of content and are ideal markup languages for the purpose of interacting with units and quantities. The extensiveness of MATHML, provided by its close to 100 intuitive (XML) structure elements for school and engineering mathematics [KR09] and multiple representation possibilities (see section 3.2), and the modularity and extensibility of OPENMATH, given by Content Dictionaries, would allow the development of great applications and services (some of which are discussed in section 4.5) that will ease the understanding of units' and quantities' semantics.

2.2 The ARXMLIV Corpus

The ARXMLIV corpus is the ideal environment for the identification of units and quantities since it contains a collection of more than 600,000 scientific publications. It is based on Cornell University's ARXIV e-Print archive [Arx] which was converted to XML for a better separation between plain text and mathematical expressions [Sta+10] and for an easier machine processable format. The whole conversion process of the ARXIV archive is based on the LaTeXML tool [Mil] for the LaTeX to XML conversion and on the ARXMLIV build system for the distributed conversion of such large document collections. It enabled the successful conversion of more than half of the scientific articles from the archive to a semantically enriched XHTML+MATHML representation which can be easily rendered by a web browser. This perspective of having scientific papers fully rendered in a web browser would allow the integration of a wide range of web services and tools, thus adding features and functionalities like mathematical formulae web search [Koh+08] and many others.

The size and broadness of the ARXMLIV corpus would enable a far greater *identifying ground* for further extension and improvement of the generic occurrence types described and exemplified in section 3.3. It will also provide a perfect testing ground for the intended implementation of the services detailed in sections 4.2 and 4.5.

The analysis will be made on top of the already existing representations of units in OPENMATH and *Content* MATHML (cf. section 2.1) with a greater focus on the latter since it is the predominant document format of the ARXMLIV corpus. The approach to identifying units and quantities in OPENMATH is also considered in section 3.1 for a broader view on the different types of representations and for the analysis of multiple perspectives such content-oriented markup languages have.

2.3 Related work

There already exists research in the field of units and quantities, especially regarding the standardization of their representation. OPENMATH has a proposed unit and dimension representation, detailed in [DN03] which also provides the current state of art described by the several Content Dictionaries (CDs) that treat units in mathematical documents. The provided in-depth analysis of the prospective representations of units and their dimensions (taking into account the pros and cons of each approach) allows for a broader view to the multitude of possibilities of capturing their semantics. Each of the treated cases is considered further in section 3.1 for a better understanding of the possibilities they offer. The extensibility of OPENMATH allowed the development of more Content Dictionaries for units; the most significant work in this direction is the research of Joseph Collins on CDs for SI (International System of Units standard [Sib]) quantities and units [Col09]. Although his work is concentrated on the creation of scientific markup for physics based models, it also provides a very strong insight on the concepts of *quantity* and *unit* and on the prospects of capturing semantics in their representation.

The advantages of OPENMATH's extensibility have also encouraged research in the direction of unit conversion services [Str08]. Besides the implementation of such a service, Jonathan Stratford's research also identifies the difficulties of unit conversion and the limitations of OPENMATH's current state of art (detailed in section 3.1) in regard to unit representation. Nevertheless his work offers an in-depth view to the representation of units and quantities in the respective content markup language.

Units and dimensions were also considered in MATHML which provides multiple encoding possibilities [HD03], one of which also integrates semantics in unit expressions (see 3.2).

Although not directly considered in this proposal, OMDOC (Open Mathematical Documents) semantic markup format for mathematical documents [Koh06] integrates the two mentioned content markup languages (at object level representation) thus providing an extensible framework for the semantic representation of units and quantities. One extension of OMDOC which goes into this direction is PHYSML (Physics Markup Language) [HKS06] which provides a *Theory of SI units* in a nutshell. This semantic markup format will be further considered in the guided research project.

3 Units in Technical Documents

This section represents a preliminary stage of the research project by providing the preparatory ground for the identification and analysis of *unit and quantity* occurrences. The first two subsections describe a detailed state of the art of *units' and quantities'* representation in OPENMATH and in *Content* MATHML (with concrete examples), while the last subsection describes the different occurrences of units and quantities in technical documents selected from the ARXIV e-Print archive [Arx].

3.1 Units and Quantities in OPENMATH

OPENMATH encompasses units through Content Dictionaries [Col09]. Although generically described in section 2.1, it is important to emphasize OPENMATH's extensibility through the creation of new content dictionaries that can add new symbols (known as OMS) with their respective semantics via Formal Mathematical Properties (known as FMPs) or simply through the extension of the already existent unit CDs.

The real problem of OPENMATH is the way units and quantities are interlinked. For example, the OPENMATH representation of a quantity (a number) would be:

```
<OMI> 100 </OMI>
```

while for units, take the unit “gramme” as example, the following representation exists:

```
<OMS name="gramme" cd="units_metric1" />
```

The representation of *100 grammes* would look as follows:

```
<OMA>
    <OMS name="times" cd="arith1" />
    <OMI> 100 </OMI>
    <OMS name="gramme" cd="units_metric1" />
</OMA>
```

which is one out of multiple representations of ***quantity*** *unit* [DN03], thus making it difficult to define a grammar that could capture all of them.

For an in-depth analysis of unit (and quantity) representation in OPENMATH, [SD08] is the ideal paper to consult.

3.2 Units and Quantities in *Content* MATHML

The characteristics of units in *Content* MATHML are extensively described in [HD03], but for the scope of this proposal only the significant ones have been listed below:

1. **Simple Units**: *“In expressing a quantity with units [...] it is recommended that the unit be the last child of an **apply** element which has the **times** element its first child”* in Section 3 of [HD03].This implies the following type of representation for a simple quantity with unit (e.g. 1.7 m):

```
<apply>
    <times/>
    <cn> 1.7 </cn>
    <csymbol> m </csymbol>
</apply>
```

2. **Compound Units**: *"It would also be best if compound units are kept separate as a nested apply at the end of a product"* in Section 3 of [HD03]. This implies the following possible representation of a quantity with simple, compound units (e.g. 350 m/sec):

```
<apply>
    <times/>
    <cn> 350 </cn>
    <apply>
       <divide/>
       <csymbol> m </csymbol>
       <csymbol> sec </csymbol>
    </apply>
</apply>
```

   Here the compound unit is represented by the division of two simple units, each enclosed in its own **csymbol** element. The use of simple units in the **csymbol** element is actually enforced throughout Section 5 of [HD03]. Semantically speaking, there exist simple units (e.g. Watt) that have a compound representation (e.g. Watt = Nm) and thus two identical expressions can have different representations.

3. **Semantic Units** (Expressions): *"In general, it is quite difficult to pick a unit from an expression. [...] a method for facilitating the transmission of unit information by making use of the* ***semantics*** *element. [...] Each* ***semantics*** *element would contain the attribute* ***definitionURL="http://.../units/"****. The first child element would be the unit encoded in either* Content *or* Presentation MATHML.". These types of expressions are detailed starting from Section 6 in [HD03]. The introduction of this attribute enables an easier spotting of units in *Content* MATHML and also facilitates the disambiguation procedure.

3.3 Common occurrence of units and quantities

For the purpose of identifying the different ways units appear in the "wild", seven papers [CB10; SGHPD10; LJ10; Stu+10; EFK10; Haz+10; Fuk+10] from the fields of Physics, Astronomy and Mathematics were randomly chosen from the ARXIV e-Print archive [Arx] and all occurrences of quantities and units were highlighted. Although the selected pool of technical documents is quite narrow, each of the selected papers contained a sufficient number of appearances in a extensive variety of contexts[1]:

- ***Inner-paragraph occurrences*** which consist of embedded unit-quantity or unit-formula blocks within sentences:
  - **simple occurrence**: simple or prefixed units together with quantities containing:
    - abbreviations of units:
      a broad temperature range $(4K - 300K)$ - [Haz+10], p. 1
      with frequencies around $1kHz$ - [SGHPD10], p. 1
      use a $500M\Omega$ *thick film resistor* - [Haz+10], p. 2
    - full names of units:
      over a hundred gigabytes *of street network data* - [LJ10], p. 2
      with a few meters *gap in between* - [LJ10], p. 2
      costs only 25 seconds *to generate* - [LJ10], p. 6
  - **complex occurrence**: compound units together with quantities (including composed quantities) or mathematical expressions:

    absorption column $N_H >= 5 \times 10^{23} cm^{-2}$ - [SGHPD10], p. 3
    cables are cut into $8 \times 1.85m$ *shorter lenghts* - [CB10], p. 6

- ***Graph axis occurrence***: occurrence of units as axis labels[2]

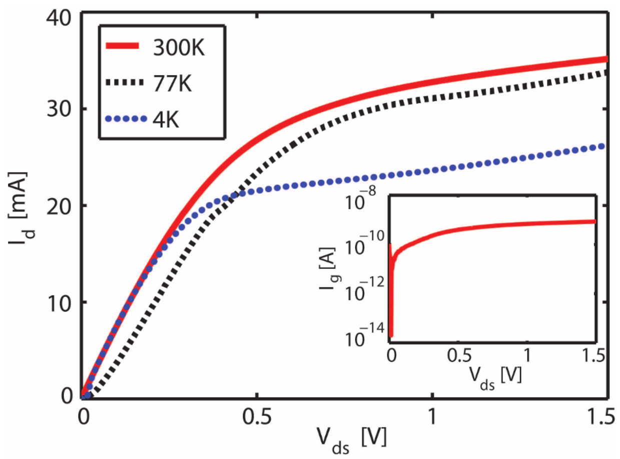

Figure 1: Output characteristics of Fujitsu FHX35X HEMT Integrated Capacitance Bridge [Haz+10]

[1]Only part of them appear in the provided examples; all the papers will be taken into account for the guided research project.

[2]Although it is quite common to have the units as axis labels for plots, it is practically impossible to interact with the data in such plots given that their representation format does not allow its reiteration. Thus this type of occurrence will be excluded from the in-depth approach in section 4.

- ***Formula / Mathematical expression occurrences*** which consist of independent formula / mathematical expression blocks linked with units:
  Such examples were not encountered in the seven selected papers mentioned at the beginning of this section but one expected occurrence would be

  $$F_{q_1 q_2} = \frac{1}{4\pi\epsilon_0}\frac{q_1 q_2}{d^2}N$$

  where N represents the unit for force in Physics ($Newtons$) and not a variable (q_1, q_2, d) or constant ($\frac{1}{4\pi\epsilon_0}$).

4 Research Approach and Goals

4.1 Analysis of Units and Measurable Quantities

Although section 3.3 mentions a wide range of context occurrences for units and quantities, when it comes to the practical aspect of identifying them, the focus should be shifted to their representation in the OPENMATH and *Content* MATHML markup languages. In OPENMATH the existence of content dictionaries for units [Col09] allows the capture of their semantics, while for *Content* MATHML the formal representation of units is extensively defined in [HD03]. A detailed description of their syntactic representation can be consulted in sections 3.1 and 3.2.

In order to provide the interaction with units and their quantities within technical documents a means of identifying them is required. The already shown common types of occurrences in section 3.3 provide a good starting point for the definition of a grammar that would encompass all the different aspects of the content markup languages that were considered (OPENMATH and *Content* MATHML). In a first stage, the identification of the described common occurrences of units and quantities will be done taking into account the defined grammar and the markup language specific syntactic rules exemplified throughout 3.1 and 3.2. This will be mainly applied to the large pool of technical documents from the ARXMLIV archive.

4.2 Implementation

Besides using the grammar rules defined for spotting units and quantities in OPENMATH and *Content* MATHML, on document level, the construction of semantic formula trees [Gin10] with the help of the LaMaPUn OOPerl library [Gin+09], which concerns natural language processing, can enable an extensive analysis of special types of occurrences that don't have a formal markup representation in OPENMATH, nor in *Content* MATHML. This would also lead to the extension of existing unit related Content Dictionaries (CDs) [Col09; SD08] to formalize some of the *exotic* occurrences of units (see section 3.3 for some examples). This identification service will enable the insertion of semantics to disambiguate the occurrences.

A multitude of services can be built on top of unit and quantity semantics (as can be seen in section 4.5). One immediate such service that can take advantage of units and quantities in technical documents is a unit converter. During the past couple of years multiple implementations of such a service have been proposed, one of the most extensive ones being [Str08]. A unit conversion service integrated into web documents is implemented within the JOBAD framework [GLR09] which has the advantage of being a web framework and thus allows high extensibility and user interaction.

4.3 Time Plan

The prospective timeline of the described guided research topic can be summarized by the following major points:

- January 2011 – *Grammar Construction for Common Unit and Quantity Occurrences in Technical Documents*
  - Further analyze the occurrence of units and quantities in technical documents
  - Define grammar rules that correspond to the identified occurrences
  - Check and document special cases
- February 2011 – *Analysis and Adaptation of Units and Quantities identified on a large pool of technical documents* - ARXMLIV *[Sta+10]*
  - Investigate existing parsing methods for *Content* MATHML on the ARXMLIV corpus
  - Start implementing a *units and quantities* identification service based on the defined grammar
  - Analyze the outcome statistics from the data obtained by running the (partially implemented) identification service on the ARXMLIV corpus
  - Use the Construction of Semantic Formula Trees [Gin10] to identify special cases
- March 2011 – *Improvement and Extension of the Scope of Identifiable Units and Quantities*
  - Finalize the implementation of the *units and quantities* identification service

  - Write Content Dictionaries for the extension of existing documented units and relations between them
  - Extend the JOBAD framework unit converter to showcase the practical part of the guided research project
- April 2011 – *Finalization of the Guided Research Thesis*

At this moment it is neither final, nor complete and unforeseen difficulties will most probably slightly modify the monthly milestones of the project. Nevertheless, the extended time period of four months (January 2011 – April 2011) would allow plenty of time for an in-depth investigation of the proposed topic.

4.4 Expected results

The goal of this guided research project is to achieve an important percentage of successful unit identification and disambiguation on the ARXMLIV corpus and to provide a framework for further extension of the ideas presented in section 4. The evaluation of the results can be done by either considering individual cases of technical documents (e.g. selecting one at random and running the service for unit identification), which would be, at first, manually verified, or by considering the entire collection of articles from ARXMLIV accompanied by a statistics service which will provide data regarding successful identifications, errors and unknown (undefined or incomplete) occurrences. These results will constitute a basis for further extension and improvement of the identification and disambiguation service (see section 4.2 for the general outline of this service).

One prospective outcome of the project would be the possibility to generate unit specific papers (e.g. having a paper with all the units in the Imperial System of Units) and also to allow unit independent search through the MathWebSearch formula search engine [Koh+08] (e.g. the search for “1 m” would also return results that have “1.0936 yards”) as well as more advanced applications and services as described in section 4.5.

4.5 Units as Enabling Technology

The prospects of unit and quantity disambiguation, as described in section 4.2 and based on 2.3, allow other relevant services/functionalities to be considered for future projects and research topics.

- **Mapping Natural Sciences Concepts to their respective Units**
  Description: defining Content Dictionaries [KR09] that would define the connection of units to general widespread natural sciences concepts like *force* (measured in Newtons: $N = \frac{kgm}{s^2}$ or any variant of the ratio) or *energy*

(measured in Joules: $J = Nm = \frac{kgm^2}{s^2} = ...$) and plenty of other examples. Only from the two basic examples one can notice the interconnection of concepts in sciences: *Energy* = *Force* $\times$ *displacement*, which can further enable scientific formula "spell checking", which might prove to be of great value to physicists, astronomers and many others.

- **Mathematical Formula Validation**
  Description: unit checking within a mathematical expression can easily allow the verification of dimensionality of the concept the expression defines. This can be further extended to a *scientific formula validation* service that can identify a wrongly written formula by matching the units of the right hand side of the equality to the units of the left hand side.

- **Unknown Unit Lookup**
  Description: as mentioned in 3.3, sometimes theoretical scientific papers only present abbreviations for concepts (e.g. N for $Newtons$, which is the unit for *force*) without mentioning anything about units/dimensions, which might turn out to be difficult for the readers who would be interested to know, for example, the order of measurement (magnitude) for the unknown measurable concepts and also a (small) description of the respective concept (e.g. Pa is the unit for *pressure*). Defining a generic way in which semantics can be added to such unknown symbols will enable showing/hiding units for expressions/formulas.

- **Unit and Quantity Semantic Search**
  Description: a service that would allow searching for any type of unit and its magnitude and return the relevant results regardless of the measuring standard of the occurrences in the paper (e.g. imperial or metric) and also regardless of their form (N or $\frac{kgm}{s^2}$).

- **Quantity and Unit's Magnitude Manipulation**
  Description: [SD08] offers an in-depth view over the representation of composed/prefixed units in OPENMATH which can capture the meaning of magnitude in a certain quantity-unit occurrence. Being able to transform for example $100N \to 0.1kN$ or $0.1 \times 10^3 N$ or $0.1 \times 10^3 \frac{kgm^2}{s^2}$ can have impressive utility when it comes to simplifying representations and adapting them consistently to a certain type of magnitude (for example *all occurrences of force expressions should have the unit represented in kN*).

As detailed at the beginning of this paper (see section 1), having a standard, uniform understanding of units and quantities can prevent hazards and even entire compatibility check processes in the industry. The presented list of prospective enabling technologies shows only a few of the endless opportunities of interacting with units and quantities in technical documents and can only be a motivation for future research in this field.

5 Conclusions

As detailed in section 1, units are sufficiently important to not be disregarded from the context of semantic documents. Unfortunately, by now, there have been only isolated approaches (see section 2.3) to solve the disambiguation issue of units and their linked quantities. Considering also the wide range of existing unit types and representations (some of which can be consulted in section 3.3), it makes it almost impossible to identify and disambiguate all of them, especially when we are talking about a collection of more than 600,000 technical documents [Sta+10]. The proposed approach (detailed in section 4) is mainly characterized by two major stages: unit and quantity identification (which can as well be divided into two identification methods through the grammar's and markup language's syntactic rules and/or through the construction of semantic formula trees) and unit and quantity disambiguation.

The challenges that are faced in the analysis of units and measurable quantities are strictly related to the semantics behind the structure of their occurrence in technical documents as detailed throughout this proposal. The batch of scientific papers used to identify the different possible occurrences of units and quantities has provided a wide range of representations (detailed in section 3.3 of this proposal).

Being able to identify and disambiguate the occurrence of units and their quantities allows a great extent of prospective applications and services (as described in section 4.5) that will enhance the interaction with units and quantities as well as their better understanding in technical documents. Considering the amount of work and detailed timeline expressed in section 4.3 it can be concluded that the subject chosen is suitable as a Guided Research topic.

References

[Air] *Aviation Safety - Air Canada Accident Report*. URL: `http://aviation-safety.net/database/record.php?id=19830723-0` (visited on 10/25/2010).

[Arx] `arXiv.org` *e-Print archive*. URL: `http://www.arxiv.org`.

[Aus+10] R. Ausbrooks et al. *Mathematical Markup Language (MathML) Version 3.0*. W3C Proposed Recommendation of 10. August 2010. World Wide Web Consortium (W3C), 2010. URL: `http://www.w3.org/TR/MathML3`.

[Bus+04] S. Buswell et al. *The Open Math Standard, Version 2.0*. Tech. rep. The OpenMath Society, 2004. URL: `http://www.openmath.org/standard/om20`.

[Car+01] D. Carlisle et al. *Conversion between MathML and OpenMath*. Tech. rep. The OpenMath Society, 2001. URL: http://www.openmath.org/documents/om-mml.pdf.

[CB10] R. T. Cahill and D. Brotherton. *Experimental Investigation of the Fresnel Drag Effect in RF Coaxial Cables*. 2010. eprint: 1009.5772v1.

[Col09] J. B. Collins. "OpenMath Content Dictionaries for SI Quantities and Units." In: *Calculemus/MKM*. Ed. by J. Carette et al. Vol. 5625. Lecture Notes in Computer Science. Springer, Sept. 2, 2009, pp. 247–262. ISBN: 978-3-642-02613-3. URL: http://dblp.uni-trier.de/db/conf/mkm/mkm2009.html#Collins09.

[DN03] J. H. Davenport and W. A. Naylor. *Units and Dimensions in OpenMath*. 2003. URL: http://www.openmath.org/documents/Units.pdf.

[EFK10] N. Evans, J. French, and K.-Y. Kim. *Holography of a Composite Inflaton*. 2010. eprint: 1009.5678v1.

[Fuk+10] K. Fukumura et al. *Modeling High-Velocity QSO Absorbers with Photoionized MHD Disk-Winds*. 2010. eprint: 1009.5644v1.

[Gin+09] D. Ginev et al. "An Architecture for Linguistic and Semantic Analysis on the arXMLiv Corpus". In: *Applications of Semantic Technologies (AST) Workshop at Informatik 2009*. 2009. URL: http://www.kwarc.info/projects/lamapun/pubs/AST09_LaMaPUn+appendix.pdf.

[Gin10] D. Ginev. *Construction of Semantic Formula Trees for ArXMLiv*. Sept. 2010.

[GLR09] J. Giceva, C. Lange, and F. Rabe. "Integrating Web Services into Active Mathematical Documents". In: *MKM/Calculemus Proceedings*. Ed. by J. Carette et al. LNAI 5625. Springer Verlag, July 2009, pp. 279–293. URL: https://svn.omdoc.org/repos/jomdoc/doc/pubs/mkm09/jobad/jobad-server.pdf.

[Haz+10] A. Hazeghi et al. *An integrated capacitance bridge for high-resolution, wide temperature range quantum capacitance measurements*. 2010. eprint: 1009.5407v1.

[HD03] D. W. Harder and S. Devitt. *Units in MathML*. World Wide Web Consortium. Nov. 2003. URL: http://www.w3.org/Math/Documents/Notes/units.xml.

[HKS06] E. Hilf, M. Kohlhase, and H. Stamerjohanns. "Capturing the Content of Physics: Systems, Observables, and Experiments". In: *Mathematical Knowledge Management (MKM)*. Ed. by J. Borwein and W. M. Farmer. LNAI 4108. Springer Verlag, 2006, pp. 165–178. URL: http://kwarc.info/kohlhase/papers/mkm06physml.pdf.

[Koh06] M. Kohlhase. OMDoc – *An open markup format for mathematical documents [Version 1.2]*. LNAI 4180. Springer Verlag, Aug. 2006. URL: `http://omdoc.org/pubs/omdoc1.2.pdf`.

[Koh+08] M. Kohlhase et al. "MathWebSearch 0.4, A Semantic Search Engine for Mathematics". manuscript. 2008. URL: `http://mathweb.org/projects/mws/pubs/mkm08.pdf`.

[KR09] M. Kohlhase and F. Rabe. "Semantics of OpenMath and MathML3". In: *22nd OpenMath Workshop*. Ed. by J. H. Davenport. July 2009. URL: `http://kwarc.info/kohlhase/submit/om09-semantics.pdf`.

[LJ10] X. Liu and B. Jiang. *Defining and Generating Axial Lines from Street Center Lines for better Understanding of Urban Morphologies*. 2010. eprint: `1009.5249v1`.

[Mar] *CNN - "NASAs metric confusion caused Mars orbiter loss"*. Sept. 1999. URL: `http://articles.cnn.com/1999-09-30/tech/9909_30_mars.metric_1_mars-orbiter-climate-orbiter-spacecraft-team?_s=PM:TECH` (visited on 10/29/2010).

[Mil] B. Miller. *LaTeXML: A LaTeX to XML Converter*. URL: `http://dlmf.nist.gov/LaTeXML/` (visited on 11/16/2010).

[Oxf] *Oxford English Dictionary. "unit" definition*. URL: `http://dictionary.oed.com/entrance.dtl` (visited on 10/29/2010).

[SD08] J. Stratford and J. H. Davenport. "Unit Knowledge Management". In: *Intelligent Computer Mathematics*. 9th International Conference, AISC, 15th Symposium, Calculemus, 7th International Conference MKM (Birmingham, UK, July 28–Aug. 1, 2008). Ed. by S. Autexier et al. LNAI 5144. Springer Verlag, 2008, pp. 382–397.

[SGHPD10] J. D. Sanabria-Gomez, J. L. Hernandez-Pastora, and F. Dubeibe. *Innermost stable circular orbits around magnetized rotating massive stars*. 2010. eprint: `1009.0320v1`.

[Sib] *The International System of Units (SI) 8. Edition*. Bureau International des Poids et Mesures, 2006. URL: `http://www.bipm.org/utils/common/pdf/si_brochure_8_en.pdf`.

[Sta+10] H. Stamerjohanns et al. "Transforming large collections of scientific publications to XML". In: *Mathematics in Computer Science* 3.3 (2010): *Special Issue on Authoring, Digitalization and Management of Mathematical Knowledge*. Ed. by S. Autexier, P. Sojka, and M. Suzuki, pp. 299–307. URL: `http://kwarc.info/kohlhase/papers/mcs10.pdf`.

[Str08] J. Stratford. *Creating an extensible Unit Converter using OpenMath as the Representation of the Semantics of the Units*. Tech. rep. 2008-02. University of Bath, June 2008. URL: `http://www.cs.bath.ac.uk/pubdb/download.php?resID=290`.

[Stu+10] M. Stumpf et al. *Resolving the L/T transition binary SDSS J2052-1609 AB * (Research Note)*. 2010. eprint: `1009.5672v1`.

[Usm] *US Metric Association "Unit Mixups" article*. URL: `http://lamar.colostate.edu/~hillger/unit-mixups.html` (visited on 10/25/2010).