Bringing Mathematics to the Web of Data: The Case of the Mathematics Subject Classification*

Christoph Lange[1,2], Patrick Ion[3,4], Anastasia Dimou[4], Charalampos Bratsas[4], Wolfram Sperber[5], Michael Kohlhase[1], and Ioannis Antoniou[4]

[1] Computer Science, Jacobs University Bremen, Germany
{ch.lange,m.kohlhase}@jacobs-university.de
[2] SFB TR/8, University of Bremen, Germany
[3] Mathematical Reviews, American Mathematical Society
ion@ams.org
[4] Web Science, Aristotle University Thessaloniki, Greece
{andimou,cbratsas,iantonio}@math.auth.gr
[5] FIZ Karlsruhe, Germany
wolfram@zentralblatt-math.org

Abstract. The Mathematics Subject Classification (MSC), maintained by the American Mathematical Society's Mathematical Reviews (MR) and FIZ Karlsruhe's Zentralblatt für Mathematik (Zbl), is a scheme for classifying publications in mathematics. While it is widely used, its traditional, idiosyncratic conceptualization and representation did not encourage wide reuse on the Web, and it made the scheme hard to maintain. We have reimplemented its current version MSC2010 as a Linked Open Dataset using SKOS, and our focus is concentrated on turning it into the new MSC authority. This paper explains the motivation and details of our design considerations and how we realized them in the implementation, presents use cases, and future applications.

1 Introduction: The MSC and Its Applications

Classification schemes – "descriptive information for an arrangement or division of objects into groups based on characteristics, which the objects have in common" [9] – are in broad use in digital libraries. Due to the long history of library sciences, existing classification schemes cover a wide range of subjects of interest. Pre-existing subject schemes have proven to be a starting point for developing corresponding formal representations of domains in the form of ontologies.

There are general purpose classification schemes such as the Dewey Decimal System (DDC; see http://www.oclc.org/dewey/ and p. 770) and the Library of

* We would like to thank Joseph Corneli from the PlanetMath encyclopedia for providing us with information on the PlanetMath access statistics. The first author has been supported by DFG project I1-[OntoSpace] of SFB/TR 8 "Spatial Cognition" and EPSRC grant "EP/J007498/1". The second author thanks the Mathematics Department of the University of Michigan for its support.

E. Simperl et al. (Eds.): ESWC 2012, LNCS 7295, pp. 763–777, 2012.

Congress Subject Headings (LCSH; cf. sec. 7), as well as domain-specific ones such as ACM's Computing Classification System (CCS [23]) and the Physics and Astronomy Classification Scheme (PACS [17]). The Mathematics Subject Classification (MSC [14]) is the most common point of reference in mathematics. It has been used to classify mathematical documents of all types, ranging from lecture notes to journal articles and books. The MSC is maintained for the mathematical community by Mathematical Reviews (henceforth abbreviated as MR) and Zentralblatt Math (henceforth abbreviated as Zbl). The MSC is a three-layer scheme using alphanumeric codes, for example: **53** is the classification for Differential Geometry, **53A** for Classical Differential Geometry, and **53A45** for Vector and Tensor Analysis.

The present version MSC2010, released in January 2010, is now in production use at MR and Zbl. Fixes of simple factual and conceptual errors are still possible, whereas larger changes will be deferred to the next major revision to guarantee a period of stability to developers of applications and services.

Current Usage All major mathematical journals and digital libraries make use of the MSC. Examples range from the services of the AMS Mathematical Reviews, online as MathSciNet, and FIZ Karlsruhe's ZBMATH[1], through almost all the publishers of mathematics (Elsevier, Springer, etc.), to the arXiv.org pre-print server and the PlanetMath free encyclopedia [18]. The MSC is mainly used as a means of structuring mathematics literature in libraries and for the purposes of retrieving information by topic. For example, a recent analysis of the PlanetMath server logs[2] shows that accesses of PlanetMath's "browse by subject" pages[3], whose structure corresponds to the MSC2000, constitute 5 to 6 percent of all accesses of PlanetMath pages. Taking, furthermore, into account that these pages are much less often linked to from external sites (such as Wikipedia) and change less frequently, one can assume that they are less frequently visited by users coming from a search engine's results pageand thus constitute a significant fraction of PlanetMath's "intra-site" traffic.

When an author, an editor, or a librarian classifies a publication, he or she typically identifies the right class(es) by consulting a human-readable version of the MSC. Web forms for creating a mathematical publication or uploading an existing one to a digital library typically require manual input of the MSC classes; the same holds for the search forms e.g. of MR or Zbl. Assistance is not provided – neither to authors, who could particularly benefit from an automatic suggestion of appropriate MSC classes based on the contents of an article, nor to users searching for articles, who could, e.g., benefit from the ability to select an MSC class without knowing its alphanumeric code, and from an automatic suggestion of related classes.

Maintenance and Revision So Far. The master source of the MSC has until recently been maintained in one plain TeX file, using a set of custom macros

[1] `http://www.ams.org/mathscinet/` and `http://www.zentralblatt-math.org/zbmath/`

[2] Personal communication with Joseph Corneli from PlanetMath, 2011-10-31.

[3] `http://planetmath.org/browse/objects/`

which have been developed around 1984, and had no major changes since then. The marked-up source of the example given above looks as follows:

```
\MajorSub 53-++\SubText Differential geometry
  \SeeFor{For differential topology, see \SbjNo 57Rxx.
  For foundational questions of differentiable manifolds, see \SbjNo 58Axx}
...
\SecndLvl 53Axx\SubText Classical differential geometry
...
\ThirdLvl 53A45\SubText Vector and tensor analysis
```

It is obvious that the TeX code is not useful for web-scale machine processing and linking. A new approach for web applications seems necessary for several reasons. Specialized subject classification schemes tend to be maintained by only a few experts in the arcane ways of the art, and the intellectual capital of a classification scheme does not necessarily become more obvious from merely reimplementing in a more standard format. However, the additional possibilities for accessing the scheme and producing different views tailored to specific audiences or purposes, which standard formats enable, may lead not only to wider adoption but even to better quality control. This is because more opportunities in distinguishing new aspects of the classification scheme arise, thus uncovering issues that may not have been identified by the few expert maintainers only.

The remainder of this section reviews the recent maintenance of the MSC implementation and points out problems we encountered. From 2006 to 2009 the MSC2000 version, then in current use, underwent a general revision, done publicly by the editorial staffs of MR and Zbl. This revision included additions and changes, and corrections of known errors and resulted in MSC2010. The editors took into consideration comments and suggestions from the mathematical public, of which there were on the order of a thousand recorded in a MySQL database. This was done using a standard installation of MediaWiki which was, and still is viewable by all, but was only editable by about 50 staff members. Each change from the previous version can be clearly seen (additions in green, deletions in red, on a yellow background)[4].

Once the intellectual content had been finalized in this process, the new MSC2010 TeX master file in the format described above had to be produced, as well as derived and ancillary documents in various formats. These included: a table of changes, a KWIC index, PDF files for printing, as well as further variant forms useful to MR and Zbl. Furthermore, a TiddlyWiki edition (a single-user wiki in one HTML file; cf. `http://www.tiddlywiki.com`) was provided to enable users to download a personal copy of the MSC2010, which they could browse and annotate. The TeX master was obtained from the MediaWiki using a custom Python script; most of the derived files were constructed from the TeX master using custom Perl scripts. Obviously, all of these scripts were specific to the custom MSC TeX format; therefore, it would neither have been possible to reuse existing scripts from the maintenance of other subject classification schemes, nor will it be possible to use our scripts for a scheme other than the MSC.

[4] see, for example, `http://msc2010.org/mscwiki/index.php?title=13-XX`

2 Requirements for a Reimplementation

The previous section outlined the current state of the MSC2010 and its limitations; an account of further problems beyond the scope of this paper can be found in [20]. To do better MR and Zbl decided to reimplement the MSC2010 after its release, i.e., after the mathematical domain knowledge had been settled. Concrete requirements and desirable goals for MSC2010 were:

1. The new implementation should **facilitate use and reuse** of the MSC:
   (a) It should give convenient access to all the capabilities in the MSC that MR and Zbl depend on in producing their services and allow development.
   (b) It should allow content providers, such as scientific publishers, to easily offer searching and querying publications using MSC classes in their digital libraries.
   (c) It should enable making tools and interfaces, and reusing existing tools, that will help authors to classify their documents in a way that can easily be processed automatically.
2. The new implementation should **facilitate maintenance** of the MSC:
   (a) It should be **complete** in that it **preserves all information** that was present in the existing implementation – preferably in a **semantically faithful** way, and leave room for **subsequent semantic refinements**.
   (b) It should be maintainable with **standard tools**, minimizing the need for custom programming.
   (c) It should enable a closer integration of maintenance-related information, e.g. changes from MSC2000 to MSC2010, which were previously recorded in separate XML files having a custom schema.
3. While the core concept scheme of the MSC is maintained through an editorial process, the new implementation should **enable knowledge workers and service developers** in mathematics and related fields to **adapt and extend** the MSC for their purposes:
   (a) To connect mathematical subjects to related subjects in other domains (e.g. science).
   (b) To enhance the MSC for their custom use cases, e.g. to add unofficial Greek class labels when using the MSC to structure a Greek lecture note repository
   (c) However, such customizations should not affect the core scheme.
4. The new implementation should not only allow defining connections to related subjects, but it should **allow end users to *explore* such connections** and **discover new relevant knowledge**.

We chose RDF Linked Data, using the W3C-standardized SKOS vocabulary (Simple Knowledge Organization System [13]), for the reimplementation – hoping that our commitment to a standard will not only facilitate *our* maintenance but also foster wide adoption. Historically, the choice was motivated when a library asked for a MARC (Machine-Readable Cataloging) version of the MSC, and experts suggested to start with SKOS, as, e.g., the Library of Congress had done for its

Subject Headings (cf. sec. 7). Moreover, we knew we could rely on existing best-practice recommendations for modeling classification systems in SKOS, such as [16].

3 Design of the MSC/SKOS Concept Scheme

This section discusses design decisions we made while we were implementing the MSC2010 in SKOS, as to satisfy as many of the given requirements as possible. We roughly divide the different aspects of the MSC by the complexity of their representation in SKOS. This section focuses on structures that could be implemented in SKOS Core in a straightforward way, or by relatively straightforward extensions, the latter being implemented as an OWL ontology whose namespace we abbreviate with *mscvocab:*. In contrast, sec. 5 discusses points where we reached the limits of SKOS, or even of RDF.

The Basic Hierarchy (SKOS Core). The MSC presents a simple tree graph with 63 first-level nodes below the top root element, and 528 nodes as children of those, with 5606 final leaves. Implementing the basic concept hierarchy required a straightforward application of the following SKOS vocabulary terms:

- *skos:ConceptScheme* – for the whole concept scheme
- *skos:Concept* – for each MSC class. In contrast, the traditional terminology explicitly represented the level of a class in the hierarchy (“Major Subject”, “Second Level”, “Third Level”), but the level is deducible by the link structure expressed by *skos:narrower*.
- *skos:hasTopConcept* – for linking the scheme to the top level of the concept hierarchy
- *skos:narrower* – for links from the top level to the second level, and from the second level to the third level
- *skos:broader* – for backlinks in the opposite direction
- *skos:inScheme* – for backlinks from each class to the scheme

The following listing shows (in Turtle serialization) the SKOS implementation of these basic properties of the MSC class 53A45, plus two properties covered by the following subsections (notation and label):

```
msc2010:53A45 a skos:Concept; skos:inScheme msc2010:; skos:broader msc2010:53Axx;
 skos:prefLabel "Vector␣and␣tensor␣analysis"@en ;
 skos:notation "53A45"^^mscsmpl:MSCNotation ;
```

The choice of appropriate URIs for the concepts required some more considerations and is therefore covered separately on p. 771.

Notations (SKOS Core) The 5-character class number (e.g. 53A45) could be represented as a notation[5] (*skos:notation*), for which, for the purpose of enabling MSC-specific validation, we implemented our own datatype *mscvocab:MSCNotation*.

[5] “a string of characters [. . .] used to uniquely identify a concept within the scope of a given concept scheme [which is] different from a lexical label in that a notation is not normally recognizable as a word or sequence of words in any natural language” [13]

Multilingual Labels (SKOS Core). About each concept, the TeX source provided as further information a descriptive English text (`\SubText` in the TeX source), which could be represented as a preferred label, except that mathematical content requires separate treatment (see p. 768). Choosing SKOS allowed us to go beyond just representing the information given in the TeX sources. Independently from the TeX source, several trusted sources had contributed translations of the descriptive texts to further languages: Chinese, Italian, and Russian[6]. SKOS, thanks to its RDF foundation, not only facilitates handling accented characters (which also occur in the English-language descriptions) and non-Latin alphabets, but allows for multilingual labels, for example:

```
msc2010:53A45 skos:prefLabel "Vector␣and␣tensor␣analysis"@en, "向量与张量分析"@zh .
```

We expect this to facilitate maintenance of the translated descriptions, as they are now part of the master source. Plus, RDF gives external developers speaking further languages the possibility to attach unofficial labels in, say, Greek, to the MSC/SKOS dataset by maintaining a separate graph containing triples such as

```
msc2010:53A45 skos:prefLabel "Διανυσματιϰή ϰαι τανυστιϰή ανάλυση"@el .
```

and then, for the desired application, merging it into the graph given by the official dataset.

Mathematical Markup in Labels (SKOS Core). The subject of mathematics involves formulas that need special markup and symbols, even within the descriptive labels of the MSC2010. Most of them, merely consisting of numbers, variable names or operator symbols, are sufficiently simple to be represented as plain Unicode text, but the semantics of mathematical expressions is often encoded in a two-dimensional layout and some labels make use of that. A detailed analysis of the TeX source shows that 215 out of 6198 labels contain mathematical markup. While the real complexity of two-dimensional markup, e.g. fractions or matrices, does not occur in these labels, and while recent Unicode versions cover most mathematical symbols (including Latin and Greek letters in various scripts such as bold or italic, sub- and superscript digits, operators and other symbols), 23 labels remain that cannot be represented in Unicode. These include: expressions in a sub-/superscript (e.g., S^{n-1} or ${}_2F_1$), non-standard sub-/superscript letters (e.g., 1^k, H^p, or v_n), sub-/superscript symbols (e.g. C^∞), and overlined operators ($\overline{\partial}$). With MathML [3], whose recent inclusion in HTML5 is expected to lead to a more widespread adoption, there is an XML markup language that is capable of expressing such layout schemata, even with fallback alternative texts for applications that do not fully support MathML. RDF supports XML literals; these are literals of datatype *rdf:XMLLiteral*. Unfortunately, this approach is not compatible with multilingual labels, for reasons we will discuss on p. 773.

```
msc2010:26E10 skos:prefLabel "<mml:math␣alttext="$C^\infty$">
␣␣␣␣<mml:msup><mml:mi>C</mml:mi><mml:mi>∞</mml:mi></mml:msup>
  </mml:math>-functions, quasi-analytic functions"^^rdf:XMLLiteral .
```

[6] The sources were: Tsinghua University for Chinese, the Russian Academy of Sciences for Russian, and Alberto Marinari for Italian (MSC2000 only)

Linked Partitively Related Concepts (Extension). In addition to links to broader and narrower concepts, the TeX source contained three further types of “see also” links from MSC concepts to other related MSC concepts. We introduced custom properties for each of these link types to capture their specific semantics. Note that these links are not symmetric; therefore, we did not make these properties subproperties of *skos:related*, but of a custom property *mscvocab:relatedPartOf*[7] (to express that it is not an overall matching but a partitive relationship), which we declared a subproperty of the generic *skos:semanticRelation*. “See also” and “See mainly” could then be represented in a straightforward way, using the custom properties *mscvocab:seeAlso* and *mscvocab:seeMainly*. Note that it was easy to identify links having MSC classes as their targets, as such targets were preceded by `\SbjNo` (“subject number”) in the TeX source. The trickier “See for” link represents conditional pointers, as can, e.g., be seen in the case of 53-++[8] in the listing on p. 765. The relationship of a source class to a target class is not universally asserted but restricted to a certain aspect of the concept. As SKOS does not offer built-in support for such links, we chose a twofold approach of (1) establishing such links unconditionally for ease of traversing (but with another *mscvocab:relatedPartOf* subproperty to avoid confusion), and (2) to reify them into resources that point to their source and their target and carry the condition as a property, to fully capture their semantics:

```
msc:53-XX a skos:Concept ;
  mscvocab:seeConditionally msc:57Rxx ;
  mscvocab:seeFor
    msc:53-XXto57Rxx-seeFor .
```

```
msc:53-XXto57Rxx-seeFor
  mscvocab:forTarget msc:57Rxx ;
  mscvocab:scope
    "for␣differential␣topology" .
```

We introduce *mscvocab:scope* as a subproperty of the SKOS annotation property *skos:scopeNote*. For now, we chose this custom approach to reification, given that (1) RDF’s built-in reification support (assigning an ID to the triple *msc:53-XX msc:seeConditional msc:57Rxx* and then stating further properties about the subject having that ID) is not recommended for use in Linked Datasets for lack of convenient SPARQL querying support [8, sec. 2.4] and is scheduled for deprecation in RDF 1.1[9], and (2) other alternatives, such as Named Graphs [5], have not yet been standardized and are therefore not yet universally supported.

Our custom approach has the disadvantage of requiring both the direct link and the reified link to be expressed redundantly. However, the effort of manually expressing this can be saved by automatically inferring the direct link from its reified representation, taking advantage of the fact that the composition of *msc:seeFor* and *msc:forTarget* implies a conditional link. In fact we do so, using rules (cf. p. 772), or alternatively the OWL 2 axiomatization $\mathit{mscvocab}{:}\mathit{seeFor} \circ \mathit{mscvocab}{:}\mathit{forTarget} \sqsubseteq \mathit{mscvocab}{:}\mathit{seeConditionally}$.

[7] Earlier SKOS versions included such properties in a “SKOS Extensions Vocabulary” (`http://www.w3.org/2004/02/skos/extensions/spec/2004-10-18.html`), which, however, has not been adopted as a standard so far.

[8] The actual MSC code is 53-XX; it is encoded as 53-++ for historical reasons.

[9] `http://www.w3.org/2011/01/rdf-wg-charter`

Linking Across MSC Versions and Other Concept Schemes (SKOS Core). While our main focus was on implementing the MSC2010 in SKOS, we also applied our TeX→SKOS translation script (see sect. 4) to the older versions MSC2000 and MSC1991. Particularly the MSC2000 is still widely in use; therefore, making explicit how closely classes match across MSC versions will aid automated migration of existing digital libraries or at least be able to assist semi-automatic migration. SKOS offers a set of different properties to express the closeness of matching across subject classification schemes. Frequently occurring cases in the MSC include concepts **unchanged** across versions (⇒ *skos:exactMatch*), **reclassifications** within an area, e.g. 05E40 "Combinatorial aspects of commutative algebra" partly replacing the MSC2000 classes 05E20 and 05E25 (⇒ *skos:relatedMatch*), and **diversification** of areas, e.g. within the area 97-XX "Mathematics education", which had 49 concepts in 2000 and 160 concepts in 2010 (⇒ *skos:broadMatch*). While we have so far only used these mapping properties across MSC versions, SKOS implementations of further subject classification schemes in related domains are to be expected soon (cf. sec. 8). In this setting, these properties can be applied analogously.

Linking to non-SKOS Concepts (Extension). The adoption of SKOS as a W3C Recommendation and the increasing awareness of the potential benefits of Linked Open Data are good reasons to expect further classification schemes to become available as SKOS datasets in the near future (cf. sec. 8). However, many relevant schemes are not currently available in a full SKOS implementation. At `http://dewey.info`, for example, there is an experimental implementation covering the top three levels of the DDC, which includes the class 510 "Mathematics" and its 8 subclasses. For the MSC, however, more fine-grained mappings to the DDC Revision 21 have already been identified. For the time being, we represent them by incorporating local placeholders for the relevant DDC concepts into our implementation, and linking to them, for example:

```
msc:53A45 skos:relatedMatch [ a skos:Concept ; dcterms:isPartOf ddc:, msc: ;
    skos:notation "515.63"^^<http://dewey.info/schema-terms/Notation> ;
    skos:prefLabel "Vector,␣Tensor,␣Spinor␣Analysis" ] .
```

In this listing, the DDC concept appears as a blank node; in any case we refrained from assigning URIs in the DDC namespace to them. While the URI scheme for the deeper levels has already been decided upon (having URIs such as `http://dewey.info/class/515.63`[10], they are not currently dereferenceable.

Collections of Concepts Besides the Main Hierarchy (SKOS Core). Some of the links within the MSC do not have single classes as their targets, but groups of classes, which do not have a common superconcept that one could instead link to. The most frequently used group of such concepts is the group of all subclasses covering historical works related to an area. In the numeric scheme, these subclasses end in -03; for instance, 53-03 is the class of historical works about differential geometry. We have grouped them as *skos:member*s of a *skos:Collection*, a

[10] `http://oclc.org/developer/documentation/dewey-web-services/using-api`

semantically weaker notion than *skos:Concept* – but that choice demands awareness of the fact that SKOS keeps collections and concepts disjoint. Similar groupings include general reference works (-00), instructional expositions (-01), and works on computational methods (-08).

```
msc:HistoricalTopics a skos:Collection ;
 skos:prefLabel "Historical␣topics"@en ;
 skos:member msc:01-XX, ..., msc:03-03, ..., msc:97-03 .
```

In addition, the MSC implicitly contains cross-area concepts such as “stability”, a property that a number of different mathematical structures may have [20]. The MSC classes related to stability are not currently systematically grouped; we just have the word “stability” to be found explicitly in the labels. Our implementation does not yet make such groupings explicit, but *skos:Collection* provides the necessary tools for doing so, after a careful conceptual analysis.

Co-Classification Policies. A further complication for modeling is introduced by the MSC specification prescribing that any resource classified with a -03 code (see above) be additionally classified with one class of the 01-XX section so as to express the specific historical aspect (e.g. “19th century” or “bibliography”). In the TeX source, the descriptive label of each -03 class contained a remark about that. While our current implementation does not yet represent that policy in a fully machine-comprehensible way, SKOS allowed us to move forward in two regards: (1) We attached the information to the collection of historical topics, i.e. in one central place, to facilitate maintenance. On p. 772 we explain how the information can be propagated to the members of the collection. (2) The *skos:note* property allows keeping the information in a dedicated place, separate from the labels of concepts.

```
msc:HistoricalTopics skos:note "Any␣resource␣classified␣as␣-03␣must␣also
␣␣␣␣be␣assigned␣at␣least␣one␣classification␣number␣from␣Section␣01." .
```

URI Syntax. Deploying a Linked Dataset requires thinking about a URI syntax [8]. In the SKOS implementation described so far, the MSC2010 dataset has around 92,000 triples (in the expanded version; see below); the RDF/XML serialization is around 7 MB large. We expect that information about few MSC-classified resources will be required in typical Linked Data scenarios, such as looking up information about an MSC-classified resource. Publications in paper-based and digital libraries are typically classified with two MSC classes; in addition to these, the superclasses may be of interest. As such applications should not be burdened with a 7 MB download, a “hash” namespace does not make sense. Conversely, applications that require full access to the MSC, such as annotation services that suggest MSC classes whose labels match a given text (as shown in fig. 1), or browser frontends to digital libraries, would rather benefit from querying a SPARQL endpoint, or their developers would preload them with a downloaded copy of the MSC dataset anyway – a possibility that is independent from the choice of namespace URI. Thus, we chose `http://msc2010.org/resources/MSC/2010/` as namespace URI for the MSC2010. For the older MSC versions, the last path components contain the respective years. The MSC-specific SKOS extension vocabulary and the MSC-specific datatype library reside in separate namespaces, as they are conceptually separate from the MSC classes and as we expect different (slower) maintenance cycles for them.

4 Deployment and Publication

We generated the new SKOS master source of the MSC2010 by a script (described below) in an iterative process while deciding on the design issues detailed above. After that, we published the data in four complementary ways, aiming to address a large audience of users and developers and enabling them to link their data to the MSC and to use the MSC in their services. All publications are available from the project homepage `http://msc2010.org/mscwork/`.

We implemented the script for the original translation of the old TEX master source of the MSC2010 to the new SKOS master source (one RDF/XML file) in Perl. It has now served its purpose and was never prepared or intended to be applicable to any other source representation of a classification scheme.

For querying the dataset, we expose it through a **SPARQL endpoint**. The **MSC Linked Wiki** frontend aims at providing easy and user-friendly navigation through the MSC. It uses SPARQL queries to the endpoint to present the classification on its pages.

We maintain **different RDF versions for different demands**. For ease of maintenance, the SKOS master source is restricted to a semantic core of RDF triples that avoids redundancy (e.g. only modeling the *skos:narrower* direction of the hierarchy). When an OWL reasoner is available, the *skos:broader* direction can be inferred automatically; however: In a Linked Data setting, where clients hop through the dataset from resource to resource in a "follow-your-nose" manner [21], it is essential to provide as many explicit links as possible. Secondly, inference support may not always be available in applications, depending on their scalability-related performance constraints.

Therefore, we have implemented an automatic expansion of the core dataset to an enriched "convenience" version, which we expose through the SPARQL endpoint and as Linked Open Data (LOD). So far, the latter is served from static RDF/XML files (one per MSC class), into which we split the one-file enriched version, but we are planning to serve the dataset through a SPARQL endpoint at `http://msc2010.org`. Additionally, we offer both versions of the dataset for download, so that application developers can import them into their triple stores. We have implemented the expansion as N3 rules. The following rule, for example, infers *skos:broader* from *skos:narrower*, effectively hard-coding the semantics of *owl:inverseOf*:

```
{ ?conc skos:narrower ?narrowerConc } => { ?narrowerConc skos:broader ?conc }.
```

Further rules infer *skos:topConceptOf* from *skos:hasTopConcept*, generate backlinks from reified "see for" links to their sources, un-reify the "see for" links into *mscvocab:seeConditionally*, dumb down all MSC-specific links to *skos:semanticRelation*, and further dumb down *skos:semanticRelation* to *rdfs:seeAlso* for off-the-shelf linked data browsers. These rules can be applied to the core dataset using an N3 reasoner such as cwm [2], which increases the number of triples by more than 16% (from 79,000 core triples to 92,000).

```
cwm --rdf msc2010-core.skos --n3 expand-skos-rules.n3 --think
```

5 Benefits Experienced and Difficulties Encountered

SKOS was designed to be simple and thus powerful in deployment. Its authors thought of it as much simpler than full OWL and very suitable for such cases of knowledge organization as thesauri. However, in the research described here, SKOS had to pass the reality check of a large classification scheme that had grown up in a specialized field over a long time. This section summarizes the benefits we experienced and the difficulties we encountered.

Relying on SKOS satisfies the requirements to "facilitate use, reuse, and maintenance" in that RDF in general and SKOS in particular enjoy wide tool support. There are tools for searching and querying, for editing, for consistency checking, and for annotating documents; for an overview, see `http://www.w3.org/2001/sw/wiki/SKOS` and [19]. Concerning reuse, the Linked Data principles provide a straightforward way of making SKOS/RDF accessible on the Web not only for browsing and for download, but also as a target for linking. Furthermore, we expect the wide availability of RDF parsers to facilitate implementing conversions into forms that are used by library management systems. In regard to maintenance, SKOS proved able to capture large parts of the structural semantics of the MSC. In particular, it supports maintaining links to other concept schemes and translations together with the core scheme.

Our implementation is complete in that it preserves all information that was present in the old TEX source – often making the semantics more explicit and thus more easily accessible to automated processing. Making the semantics explicit was partly supported by SKOS itself, but most of it had to be done by extensions – mostly using extension points SKOS or RDF provided for.

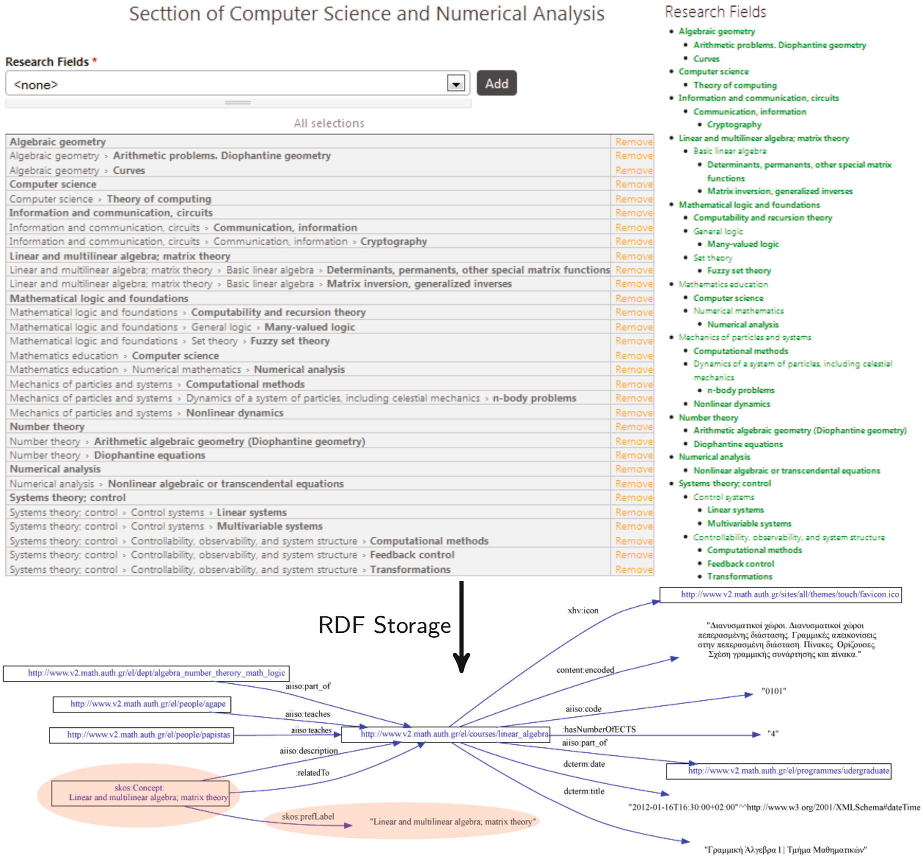
Multilingual Labels vs. Mathematical Markup. One particular problem remains, for which neither SKOS nor RDF provided a sufficient solution. On p. 768 we pointed out the importance of formulas in labels in the mathematical domain. XML literals using MathML seem to be the solution, but for the following reasons they conflict with multilingual labels, which are also highly relevant in the MSC setting: (1) The SKOS recommendation states that "by convention, skos:prefLabel [is] only used with plain literals" [13, sec. 6.5.4], i.e. with non-datatyped literals. (2) Datatyped literals may not carry a language tag. The language of an XML literal may be indicated *inside* the XML, e.g. by enclosing the whole literal into an element that carries an *XML* language tag (e.g. `<element xml:lang="en">`) – but these are are not part of the RDF graph and therefore not accessible from SPARQL queries. (3) An English and a Greek *skos:prefLabel* with mathematical markup for the same MSC class, marked up as shown in (2), would count as two *skos:prefLabel*s without an (RDF) language tag. While that would not explicitly violate SKOS integrity condition S14, which demands that "a resource has no more than one value of *skos:prefLabel* per *language tag*", it would contradict convention (1), leaving little hope for tool support. Carroll and Philipps proposed an extension to the RDF semantics in [4] that would allow for indicating the language of an XML literal in 2005, but that idea has never been adopted. Therefore, our current SKOS implementation of the MSC2010 leaves this problem unsolved for now. Note that separating the mathematical expressions in labels from the surrounding text would not qualify as a workaround, as (1) expressions can be scattered over multiple places in a text sentence, e.g. in the role of an adjective qualifying a noun, and as (2) the structure and presentation of mathematical expressions may vary depending on the language – not in the concrete case of the MSC labels but in general.

6 Use Case: The Linked Universities Initiative

This section presents a use case for the Linked Data implementation of the MSC, highlighting its relevance for electronic publishing and education. *Linked Universities* (`http://linkeduniversities.org`) is an alliance of European universities engaged in exposing their public data as linked data. The School of Mathematics at Aristotle University Thessaloniki (AUTH), in conjunction with "Semantic AUTH", AUTH's contribution to the Linked Universities initiative, has one such semantic portal at `http://www.math.auth.gr`. The courses offered in the school are semantically annotated using appropriate ontologies such as the Academic Institution Internal Structure Ontology (AIISO), Bowlogna and Bibliographic Ontology (BIBLIO), and published according to the Linked Data principles. Furthermore, the scientific fields covered by courses, as well as the faculty's research interests, are annotated using MSC/SKOS. They will also include references to other Linked Data entities inside or outside the website.

7 Related Work

Besides following the best practices established for SKOSifying the DDC [16], our work was inspired from the LCSH dataset [22]. Both are comprehensive, general-purpose classification schemes in contrast to the domain-specific MSC. The LCSH was converted from a MARCXML representation to SKOS, using custom scripts similar to ours. Similar to our approach, the authors developed custom extensions to SKOS (e.g. structured change descriptions similar Panzer's and Zeng's, which we reused), and finally evolved them into the MADS/RDF data model, which can be thought of a superset of SKOS "designed specifically to support authority data as used by and needed in the LIS [library and information science] community and its technology systems" [12]. They also experienced limitations of SKOS, concretely concerning the representation of "pre-coordinated concepts", i.e. subject headings combined from other headings. While our Web



Secttion of Computer Science and Numerical Analysis

Research Fields *

<none> Add

All selections

Algebraic geometry	Remove
Algebraic geometry › **Arithmetic problems. Diophantine geometry**	Remove
Algebraic geometry › **Curves**	Remove
Computer science	Remove
Computer science › **Theory of computing**	Remove
Information and communication, circuits	Remove
Information and communication, circuits › **Communication, information**	Remove
Information and communication, circuits › Communication, information › **Cryptography**	Remove
Linear and multilinear algebra; matrix theory	Remove
Linear and multilinear algebra; matrix theory › Basic linear algebra › **Determinants, permanents, other special matrix functions**	Remove
Linear and multilinear algebra; matrix theory › Basic linear algebra › **Matrix inversion, generalized inverses**	Remove
Mathematical logic and foundations	Remove
Mathematical logic and foundations › **Computability and recursion theory**	Remove
Mathematical logic and foundations › General logic › **Many-valued logic**	Remove
Mathematical logic and foundations › Set theory › **Fuzzy set theory**	Remove
Mathematics education › **Computer science**	Remove
Mathematics education › Numerical mathematics › **Numerical analysis**	Remove
Mechanics of particles and systems › **Computational methods**	Remove
Mechanics of particles and systems › Dynamics of a system of particles, including celestial mechanics › **n-body problems**	Remove
Mechanics of particles and systems › **Nonlinear dynamics**	Remove
Number theory	Remove
Number theory › **Arithmetic algebraic geometry (Diophantine geometry)**	Remove
Number theory › **Diophantine equations**	Remove
Numerical analysis	Remove
Numerical analysis › **Nonlinear algebraic or transcendental equations**	Remove
Systems theory; control	Remove
Systems theory; control › Control systems › **Linear systems**	Remove
Systems theory; control › Control systems › **Multivariable systems**	Remove
Systems theory; control › Controllability, observability, and system structure › **Computational methods**	Remove
Systems theory; control › Controllability, observability, and system structure › **Feedback control**	Remove
Systems theory; control › Controllability, observability, and system structure › **Transformations**	Remove

Research Fields

- **Algebraic geometry**
  - **Arithmetic problems. Diophantine geometry**
  - **Curves**
- **Computer science**
  - **Theory of computing**
- **Information and communication, circuits**
  - **Communication, information**
    - **Cryptography**
- **Linear and multilinear algebra; matrix theory**
  - Basic linear algebra
    - **Determinants, permanents, other special matrix functions**
    - **Matrix inversion, generalized inverses**
- **Mathematical logic and foundations**
  - **Computability and recursion theory**
  - General logic
    - **Many-valued logic**
  - Set theory
    - **Fuzzy set theory**
- Mathematics education
  - **Computer science**
  - Numerical mathematics
    - **Numerical analysis**
- Mechanics of particles and systems
  - **Computational methods**
  - Dynamics of a system of particles, including celestial mechanics
    - **n-body problems**
  - **Nonlinear dynamics**
- **Number theory**
  - **Arithmetic algebraic geometry (Diophantine geometry)**
  - **Diophantine equations**
- **Numerical analysis**
  - **Nonlinear algebraic or transcendental equations**
- **Systems theory; control**
  - Control systems
    - **Linear systems**
    - **Multivariable systems**
  - Controllability, observability, and system structure
    - **Computational methods**
    - **Feedback control**
    - **Transformations**

RDF Storage
http://www.v2.math.auth.gr/el/dept/algebra_number_therory_math_logic
aiiso:part_of
http://www.v2.math.auth.gr/el/people/agape
aiiso:teaches
http://www.v2.math.auth.gr/el/people/papistas
aiiso:teaches
aiiso:description
:relatedTo
skos:Concept:
Linear and multilinear algebra; matrix theory
skos:prefLabel
"Linear and multilinear algebra; matrix theory"
http://www.v2.math.auth.gr/el/courses/linear_algebra
xhv:icon
http://www.v2.math.auth.gr/sites/all/themes/touch/favicon.ico
content:encoded
"Διανυσματικοί χώροι. Διανυσματικοί χώροι πεπερασμένης διάστασης. Γραμμικές απεικονίσεις στην πεπερασμένη διάσταση. Πίνακες. Ορίζουσες. Σχέση γραμμικής συνάρτησης και πίνακα."
aiiso:code
"0101"
:hasNumberOfECTS
"4"
aiiso:part_of
http://www.v2.math.auth.gr/el/programmes/udergraduate
dcterm:date
"2012-01-16T16:30:00+02:00"^^http://www.w3.org/2001/XMLSchema#dateTime
dcterm:title
"Γραμμική Άλγεβρα I | Τμήμα Μαθηματικών"

Fig. 1. Annotating the scientific fields of a course on the new AUTH School of Mathematics site, using MSC/SKOS and Drupal 7 semantic mappings (cf. [6])

frontend to the MSC is in an early stage, the LCSH dataset is served via a comprehensive frontend that offers each record for download in different formats (including full MADS/RDF vs. plain SKOS), a graph visualization, and a form for reporting errors. Limitations of SKOS and the possibility of extending SKOS have also been reported for domain-specific classification schemes; see, e.g., van Assem's case studies with three different thesauri [1]. In domains closely related to mathematics, we are not aware of completed SKOS implementations, but of work in progress, for example on the ACM CCS [23] for computer science[11].

8 Conclusion; Roadmap Towards a Math. Web of Data

With this work we delivered the first complete LOD implementation of the MSC. This brought us closer to satisfying our original requirements, and the rigorous

[11] Personal comm. with Bernard Rous, ACM Director of Publications, 2011-06-08.

conceptual modeling approach helped to uncover new issues in the MSC conceptualization. While this paper focuses on preserving all information from the previous TEX master sources (plus translated labels), we have also, in previous work [20], identified directions for *enhancing* the conceptual model by precise *definitions* of the MSC classes, adding index terms to classes (for which Panzer and Zeng provide a SKOS design pattern [16]) and a *faceted structure* (which the collections introduced on p. 770 only partly address).

The MSC/SKOS dataset is also one of the first Linked Datasets in mathematics. Our previous work has laid the conceptual and technical foundations for integrating mathematics into the Web of Data [11]; we believe that the availability of the central classification scheme of this domain as LOD will encourage further progress. Deploying the MSC as LOD makes it more easily reusable and enables classification of smaller resources of mathematical knowledge (e.g. blog posts, or figures or formulas in larger publications), instead of the traditional approach of assigning few MSC classes to a whole article. For a closer integration of mathematical resources with those from related domains, we plan to establish links from and to the ACM CCS [23], once available in SKOS, and with the PACS [17], which we expect to reimplement ourselves. As further deployment targets, we envision the European Digital Math Library [7], whose developers are starting to work on Linked Data publishing, as well as the PlanetMath encyclopedia, which is being reimplemented using the Planetary social semantic web portal [10]. With this deployment strategy and the increased ability to classify fine-grained mathematical resources over the Web, we also believe that the MSC/SKOS dataset may support a democratization of scientific publishing, and, by taking away some of the control from the big publishing companies and giving it back to the authors, encourage the rise of networked science that depends on collaborative intelligences [15].

References

[1] van Assem, M.F.J.: Converting and Integrating Vocabularies for the Semantic Web. PhD thesis, Vrije Universiteit Amsterdam (2010), http://hdl.handle.net/1871/16148

[2] Berners-Lee, T.: Cwm. A general purpose data processor for the semantic web (2009), http://www.w3.org/2000/10/swap/doc/cwm.html

[3] Mathematical Markup Language (MathML) 3.0. W3C Recommendation (2010), http://www.w3.org/TR/MathML3

[4] Carroll, J.J., Phillips, A.: Multilingual RDF and OWL. In: Gómez-Pérez, A., Euzenat, J. (eds.) ESWC 2005. LNCS, vol. 3532, pp. 108–122. Springer, Heidelberg (2005)

[5] Carroll, J.J., et al.: Named Graphs, Provenance and Trust. In: WWW, pp. 613–622. ACM (2005)

[6] Corlosquet, S., Delbru, R., Clark, T., Polleres, A., Decker, S.: Produce and Consume Linked Data with Drupal! In: Bernstein, A., Karger, D.R., Heath, T., Feigenbaum, L., Maynard, D., Motta, E., Thirunarayan, K. (eds.) ISWC 2009. LNCS, vol. 5823, pp. 763–778. Springer, Heidelberg (2009)

[7] EuDML – European Digital Mathematics Library, http://eudml.eu

[8] Heath, T., Bizer, C.: Linked Data: Evolving the Web into a Global Data Space. Morgan & Claypool (2011), http://linkeddatabook.com

[9] Information technology – Metadata registries (MDR) – Part 1: Framework. Tech. Rep. 11179-1, ISO/IEC (2004)

[10] Kohlhase, M., et al.: The Planetary System: Web 3.0 & Active Documents for STEM. Procedia Computer Science 4, 598–607 (2011), https://svn.mathweb.org/repos/planetary/doc/epc11/paper.pdf

[11] Lange, C.: Enabling Collaboration on Semiformal Mathematical Knowledge by Semantic Web Integration. Studies on the Semantic Web, vol. 11. IOS Press, Amsterdam (2011)

[12] MADS/RDF primer. Status: Final Public Review Document (2011), http://www.loc.gov/standards/mads/rdf/

[13] Miles, A., Bechhofer, S.: SKOS Simple Knowledge Organization System Reference. W3C Recommendation (2009), http://www.w3.org/TR/skos-reference

[14] MSC 2010 (2010), http://msc2010.org

[15] Nielsen, M.: Reinventing Discovery: The New Era of Networked Science. Princeton University Press (2011)

[16] Panzer, M., Zeng, M.L.: Modeling Classification Systems in SKOS: Some Challenges and Best-Practice Recommendations. In: International Conference on Dublin Core and Metadata Applications (2009), http://dcpapers.dublincore.org/index.php/pubs/article/view/974/0

[17] Physics and Astronomy Classification Scheme, PACS (2010), http://aip.org/pacs/

[18] PlanetMath.org, http://planetmath.org

[19] Solomou, G., Papatheodorou, T.: The Use of SKOS Vocabularies in Digital Repositories: The DSpace Case. In: Semantic Computing (ICSC), pp. 542–547. IEEE (2010)

[20] Sperber, W., Ion, P.: Content analysis and classification in mathematics. In: Classification & Ontology, Intern. UDC Seminar, pp. 129–144

[21] Summers, E.: Following your nose to the Web of Data. Information Standards Quarterly 20(1) (2008), http://inkdroid.org/journal/following-your-nose-to-the-web-of-data

[22] Summers, E., et al.: LCSH, SKOS and Linked Data. In: Dublin Core (2008), arXiv:0805.2855v3 [cs.DL]

[23] The 1998 ACM Computing Classification System (1998), http://www.acm.org/about/class/ccs98