

Towards CoPing with Information Overload

Christine Müller

Computer Science, Jacobs University, Bremen, Germany

c.mueller@jacobs-university.de

Abstract

The problem of *information overload* has been addressed by several systems. However, many approaches are limited to *informal* artifacts and need to refer to the user for reference on the quality or usefulness of retrieved information. We make use of *semantic technologies*, which facilitate the *reification* and *extraction* of *scientific practice*. Based on *semantic differences* and *similarities* of *semantically marked up* artifacts, we identify *clusters* of users with *shared practice*, i.e. *virtual communities of practice*. The *common context and preferences* of these communities can help users to *cope* with the selection, structuring, and adaptation of information.

1 Motivation

The rise of modern technology has resulted in numerous specialized tools that support various scientific activities, but none of these tools provides an all-surrounding functionality. An *all-embracing implementation even seems impossible* since the requirements of scientists are very *diverse* and even *contrary*. In particular, the choice of tools often depends on the scientist's basic assumptions and foundations, which depend on his "personal preferences and the character of the current problem" [Rab08]. Consequently, efforts are made to *integrate existing scientific tools* and their *corpora of scientific artifacts*¹.

But integration is only one side of the story, since scientists also need support to *cope* with the explosive growth of scientific information available online: They have to *select relevant content*, *structure* it, and potentially want to *adapt* it into *convenient presentations*. These activities can be influenced by a number of factors: For example, the *current problem*, the *user's individual preferences*, the *area of application*, *national conventions*, the *level of sophistication*, the *audience*, and the *historical period* are potential criteria. However, the problem of *information overload* is not new, but has been addressed by several contributions across various research areas such as *information retrieval*, *bibliometry*, *knowledge management*, *eLearning* (cf. Section 3). However, many approaches are limited to *informal* artifacts and need to refer to the user for reference on the quality or usefulness of retrieved information.

We believe that *semantic technologies* facilitate the *reification* and *extraction* of *scientific practice*, which is inscribed into scientific documents (cf. Section 4 and 5).

¹Please note that the integration of specialized scientific tools is still highly challenging and visionary. For further information for the course of our work see [Rab08; Koh06; MK08a].

Based on the markup of practice, we compute semantic differences and similarities between scientists, which eventually facilitates the clustering of users with shared practice, i.e. to compute *virtual Communities of Practice* (cf. Section 2). These virtual communities provide *common context and preferences*, which potentially facilitate a more semantic and context-aware *selection*, *structuring*, and *adaptation* of information and, thus, means to *cope* with the increasing amount of information (cf. Section 6).

2 Scientific Communities of Practice

In the late 80s [LW91] coined the term *Communities of Practice (CoP)* to express the need for a new theory of learning. Nowadays, the concept is a well-known and widely accepted theory, which has a great impact on various disciplines: Meant to be useful for the debate on education, the concept has been applied to domains such as government, science, as well as industry and is of interest to both, researchers and practitioners.

We apply the theory of CoPs to the *Science, Technology, Engineering, and Mathematics (STEM)* disciplines. We view *STEMicians* as mathematical practitioners, who understand mathematics as the *language of science* and as the *basis for several disciplines*. The STEM and other scientific community are more heterogeneous than their industrial counterparts and join researchers with different specialities and background (cf. [KW05]). Moreover, deepening knowledge and learning takes place as scientists participate in various communities, while frequently "switching the role of novice and expert depending on the current situation" [KW05]. We believe that CoPs provide *common context and preferences*, that help their members to cope with different communities' repertoires, which is particularly helpful for *novice* and *new members*. Moreover, as CoPs act as "platforms for building a reputation" [Wen05], they provide a notion of *trustworthiness*, *relevance*, and *quality* on which *less-experienced* scientists can build on.

We observe that *STEMicians primarily interact via their artifacts* including documents in a more traditional understanding such as conference proceedings, journal papers, and books as well as documents in a wider interpretation such as forum postings, ratings, and tags. We assume that scientific interactions, and more generally mathematical practice, are *inscribed* into artifacts and aim at *extracting* the *inscribed scientific practice* to *model* scientific communities and their *common preferences*.

3 State of the Art for CoP Modeling

In the following, we list research that addresses the information overload and adaptation of information wrt. to selecting, structuring, and presenting artifacts. Most of these

approach focus on the computation of *informal documents* using statistical techniques such as *citation*, *co-word*, and *keyword analysis* to model users, *social relations*, or construct *social networks* for a *personalized* and *context-aware* information access.

[CC08] view *scientific activities* as complex process that involves *heterogeneous actors*, who collaborate and publish articles that “synthesize a state of knowledge at a given time” [CC08]. Scientific publications are thus *products* of scientific communities and the main *communication medium* for scientists. [CC08] emphasize that publications and, particularly, the *interrelation* of their *inscribed concepts* (terms) *form* the *scientific landscape* or *structure*. Building on Kuhn’s notion of *paradigms* [Kuh96], the authors assume a strong correlation between the structure of science and the structure of *terms co-occurrence* across the massive collection of online publications. As a case study the authors aim at reconstructing and representing the evolution of the *complex system community*. [CC08] contribute to a more intelligent scientific database management by facilitating the *browsing of relevant articles* based on *keyword sets*. The authors relate their work to the area of *scientometric research*, i.e. the study of science or technology based on quantitative data and point out two main methods: *Citation-based analysis* and *co-word analysis*, whereas they build on the latter.

Citation analysis is based on *bibliometric coupling*, building on the frequency with which two documents are cited together, or *bibliography coupling*, based on the shared set of references of two documents. For example, [KW05] model the participation of the *Computer Supported Collaborative Learning (CSCL)* community by applying a *citation analysis* on the CSCL conference proceedings. The authors use citation analysis as *social network approach*, i.e. as representation of social relation among CSCL scientists. Considering data such as name, country and continent, discipline, conference, co-authors, and referenced authors, their model focuses on the *international distribution* and *continuity* of authors, participants, and program committee to *describe how stable and global* the CSCL community is.

[KM07] describe the *computation of social networks* for *document sharing* among *like-minded people*. Each user organizes his documents in a *collection of folders* using his own *topic hierarchy* or *classification schema*. For each folder, the user’s *personal software agent* provides a *summary*, i.e. a *keyword-list* based on the *description* of the folder’s documents. In order to *identify relevant documents* wrt. to that summary, the agent applies a *community formation algorithm* and sends its request to all members (agents) of the *computed community*. The respective agent then returns all documents in the highly correlated folders of its user’s folder collections. The approach was applied to a *collaborative bookmark management system*, in which documents are represented by bookmarks, i.e. a tuple of the document’s ID and a set of keywords describing the documents. The similarity of documents is based on the similarity of their bookmarks, which is defined as a weighted sum of two basic similarities defined over URLs and keyword lists. Recommendations are *approved by the users* providing a feedback loop to allow agents to learn.

[ZYAQ07] propose seeking expertise and information along *social network*, which “has proven a more personalized, context-based, interactive hence more efficient way to accomplish the task [of information seeking] compared to a web or formal document search” [ZYAQ07]. The *Information via Social Network (SISN)* Java toolkit is describe,

which integrates multiple communication channels, such as email or Instant Messages (IM). SISN extracts social networks based on communication patterns of users. The resulting social networks are referred to as *ego-networks*, in which the user is located at the center and his social contacts are directly linked to him. These ego-network are connected via certain distance and similarity calibration, resulting in peer-to-peer topologies along which search for expertise and information can be propagated. To build user profiles, SISN generates *keyword vector* by indexing different types of documents such as PDF, Word, HTML, bookmarks, and emails taking privacy issues into account. These vectors are used to categorize documents into different categories, which combined constitute the user profiles, also referred to as *category level profile*.

Content-Based Recommender System [AT05] recommend items similar to the ones users preferred in the past. They work well on text documents and are mostly measuring the frequency of specific keywords appearing in documents. However, these system cannot distinguish between *well written* and *badly written* text; require *prior investments* to initialize user models; and neglect items that do not match against the user profiles. *Collaborative Recommender Systems* [AT05] recommended items that people with similar tastes and preferences liked in the past and measure the similarity of relations between users instead of keyword frequencies in documents. Thus, they can be applied to any kind of content, even to items that are dissimilar to those known by the user. However, the initialization problem remains.

All approaches above focus on informal documents to extract social relations and behavior. We believe that semantic technique allow a more reliable markup and extraction of scientific practice such as the user’s basic assumption and background knowledge or his notation practice. We refer to [KK08], who provide a detailed discussion on semantic knowledge management and point out that although powerful software system for document management exists, they “cannot *interpret the documents* on the web and therefore cannot support knowledge work at a web scale” [KK08]. The authors call for *intelligent content*, i.e. “semantically enhanced learning objects and active documents that carry machine-interpretable unambiguous accounts of their meaning” [KK08] and claim that “only if computers can understand semantics, then data can become reified knowledge” [KK08].

For example, the eLearning system ACTIVE-MATH [Act07] makes use of semantic technologies to model the user’s background and knowledge to generate user-specific courses. These *learner models* [Mel01] include concepts, technically pointers to a collection of educational semantically marked up artifact in our XML-based Open Mathematical Document Format (OM-DOC) [Koh06], as well as competencies and are used to *select* and *structure* appropriate course fragments to compile individualized study material. However, user model have to be *initialized* which requires *prior interaction* and *manual specification* of preferences.

[WM07] propose *CoP support* in the interactive mathematical mediator PLATΩ. The authors follow a *document-centered* approach and consider the *notation contexts* of documents as a *dynamic parameter* separate from the document’s content. By processing the marked up notations in a document, the PLATΩ system is able to *extract the semantics* of the mathematical notations contained in the document and, thus, can model the author’s *notation practice*. Moreover, the document can be automatically adapted in

the case of notational changes to the author’s preferences. The individual practice of authors is compared and used to identify CoPs that share specific notation preference. Once having identified CoPs, PLATΩ can actively support the community members e.g. by suggestion the CoP’s standard notation, by notifying about conflicts or even by translating documents between CoPs.

4 The Semantics of Documents

As we call for the markup of artifacts for more intelligent and machine-processable content, we first need to understand the semantics of these artifacts, in particular, documents. In the following we discuss different *document layers*, which introduce our notion of documents and their semantic [Koh06; KMM07b; KK08]: The *content layer* consists of knowledge-centered fragments that contain the knowledge conveyed in the artifact. In contrast, the *narrative layer* is used to reference these content fragments and add the narrative and didactic structure of the artifact. The *presentation layer* is used to specify presentation-specific contexts such as the preferred notation system [KMR08], layout, or output format.

In addition to the three document layers, we introduce a *social layer*, which encodes the social relation between users and, thus, represent their social network. Moreover, we introduce a *meta layer*, which pervades all four layers and includes artifacts that define the social, presentational, structural, and content relations.

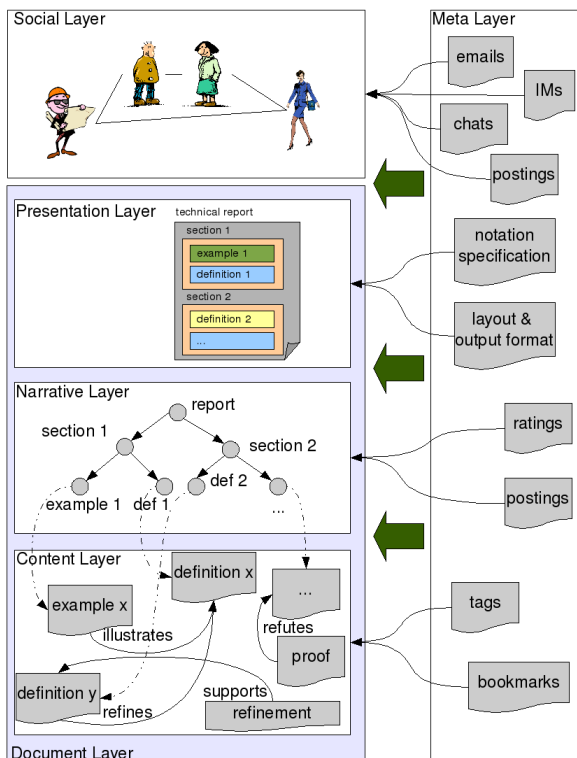


Figure 1: Document layers

Figure 1 illustrates the layers of a technical document. On the *content layer*, we distinguish *knowledge concepts* such as *example*, *definition*, *proof*, or *refinement*. These concepts are interrelated via ontological relations such as *refutes*, *supports*, *illustrates*, or *refines*. On the *narrative layer* we distinguish structural concepts such as *report*, *section*, *subsection*, and *paragraph* which reference the fragments on the content layer and provide a narrative structure. We can think of a tree, whereas the nodes represent the narrative structure, and the leaves point to fragments in the content layer. For example, *section 1* includes

two child-nodes, namely *example 1* and *definition 1*, which point to the fragments *example x* and *definition x*, respectively. *Cross edges* in the tree are relations on the narrative layer, such as *citations* within and across documents. On the presentation layer, a specific *output format*, e.g. PDF, L^AT_EX, or HTML, *layout settings*, such as textwidth, color, or fonts, and *mathematical notations* are selected. The social layer models the relations between users by extracting information from the artifacts: For example, relations, such as *knows*, *trusts*, or *collaborates with*, can be based on user roles such as *author*, *coauthors*, *referenced authors*, and *readers*. On the meta layer, we find various artifacts that pervade all layers. For example, *notation*, *output*, and *layout specification* influence the presentation of the report; while *ratings*, *postings*, *bookmarks*, and *tags* can be used to select relevant and trustworthy content as well as to structure it, respectively. In contrast, artifacts such as *emails*, *chats*, *postings*, or *IMs* define social relations.

The document and meta layer form the *artifact layer*. The relations between social layer and artifact layer, such as *writes*, *reads*, *implements*, *watches*, or *subscribes to*, allow the *bidirectional propagation of information* between the layers, that is *bottom up* and *top down*: By analysing the content, structural, presentation, and *metadata* of a document, we can identify similarities that eventually propagate to the social layer, i.e. the artifacts interrelation can be used to *construct social networks* or to *predict similarities between users*. In contrast, a top down approach allows to use the information on users and their social relation in order to *define* the adaptation, selection, and structuring of artifacts.

The bottom-up approach is used in the previous introduced work of [CC08], who use a co-word analysis to reconstruct the structure of the complex system community; [KW05], who use citation analysis to describe the CACL community; [KM07], who compute communities based on keyword-description of documents, or [ZYAQ07], who extract social networks from emails and IMs.

The top-down approach refers to user or group modeling approach such as demonstrated by the ACTIVEMATH group, who base the generation of courses on user models, which point to the artifact layer as well as collaborative recommender system, which based recommendations on the similarity of relations between users. The top down approach also refers to the wide area of Web2.0 technologies, such as social bookmarking or tagging.

A *combination* of bottom up and top down approach has been proposed in [WM07]: Semantic markup of notation facilitates the modeling of users and the computation of CoPs (bottom up). These social structures are then used to provide services on the artifact layer, such as a consistent use of notations, resolution of conflicts, or the translation of notations (top down). Alternatively, [CC08] emphasize that their approach can be used for keyword-based browsing of relevant articles; [KM07] provide recommendation and document sharing based on the computed communities; and [ZYAQ07] provide an infrastructure for seeking expertise and information along the constructed social networks. However, none of these approaches makes use of the *full power of semantic markup*.

5 The Power of Semantic Markup

Instead of focusing on informal documents as most approaches in Section 3, we build on our *semi-formal* seman-

tic representation format OMDOC². In the following we point out the strength of semantic markup:

5.1 Semantics & Context

[KK08] introduce the *semantics* of a *knowledge object*³ as “determined by its *structure* (how is the object built up from already known objects, how is it defined in terms of other objects) and its *context* (what do we already know about these objects, how are these objects defined, what is their relations to other objects)” [KK08]. The authors explore contexts that build on the content as well as on the learner. Our notion of *adaptation*, is denoted by the *recontextualization* of artifacts, which depends on the *context dependencies* of content as well as the human *ability of accommodation*, i.e. the ability to *adjust to new circumstances*. In contrast, previously mentioned approaches in Section 3 are limited in that they only refer to term occurrence and frequency and cannot compute dependencies or representations of human abilities.

In this sense, semantics provides a better notion of *trustworthiness*, *relevance*, *usefulness*, and *quality*. This improves existing approaches, which base recommendation on keywords or citations and cannot distinguish between e.g. *well written* and *badly written* documents or *relevant* and *useless* information. For example, semantic approaches elaborate citation-based recommendation by the mark up of (content) relations and can distinguish whether a reference signifies the *refutation*, *refinement*, or *support* of other artifact.

5.2 Reification of Practice

Semantic technologies enhance informal information towards meaningful and machine-interpretable representations, which facilitate the reification of scientific practice. In the following, we illustrate how we use OMDOC to reify scientific practice, such as the background knowledge and basic assumption of scientists, their choice of motivating examples, or notation preferences:

OMDOC distinguishes four level of markup: the *object level*, the *statement level*, the *theory level*, and the *document level* (cf. [Koh06]). Mathematical objects, such as functions, groups, or differential equations, are represented by markup formats such as MATHML [W3C03] or OPENMATH [Ope07], which are integrated on the OMDOC object level. Since mathematical practitioners use e.g. definitions, theorems, lemmas, or proofs as main communication means, these statements are represented on the OMDOC statement level. The *large-scale structure* and *context* in mathematics can be found in *networks of mathematical theories*, which are marked up on the OMDOC theory level. The OMDOC document level provides the markup of the content, narrative, and presentation layer of documents; whereas the content layer subsume the three previously mentioned markup levels.

Other (mathematical) markup language provide *fully formalized* artifacts, which can be computed by formal systems such as mathematical libraries and theorem provers. However, full formalization is *tedious* work and *not appropriate for capturing scientific practice*. In contrast, OMDOC provides means to annotate the *structural semantics*

²Please note that we are not restricted to the OMDOC format but emphasize on the OMDOC functionality. This approach can be applied to other powerful markup formats, such as \LaTeX [Koh05] or CNXML [HG07].

³We view knowledge objects as “tangible/visual information fragment potentially adequate for reuse, which constitute the content of documents” [Mül06]

of artifacts, that is “the structure, the meaning of text fragments, and their relations to other knowledge” [Koh06] on *all levels and layers*. This markup facilitates the automatic processing of scientific artifacts, while allowing authors to choose the level of formality: Formal fragments, programming code as well as informal but human-oriented formats can be included, promoting OMDOC towards a *hybrid* format. Consequently, OMDOC does not require the full formalization of content and can be used to *reify scientific practice*.

Reifying Background and Basic Assumption

The markup on theory level provides the reification of a scientist’s *interest*, *background*, *focus*, or *basic assumptions* and, thus, elaborates *expertise models* based on document metadata. In the context of CoPs, *common pointers* or use of OMDOC theories facilitate the computation of *similarities* among scientists and eventually the *identification of CoPs* with shared *interest* and *domains*. Figure 2 illustrates a statement and theory level in which a *definition y* points to *theory x*; while *definition x* points to *theory xx*. Assuming that both definitions are used in separate documents of user *A* and *B*; a *similarity measure* based on the *theory pointers* would define them to be unequal. However, we are able to relate theories, e.g. via (*iso*)*morphism* and *logic translation* [Rab08]. Given that a *mapping* from *theory x* to *theory xx* can be identified, both users would be equal with respect to their *common theoretical assumptions*.

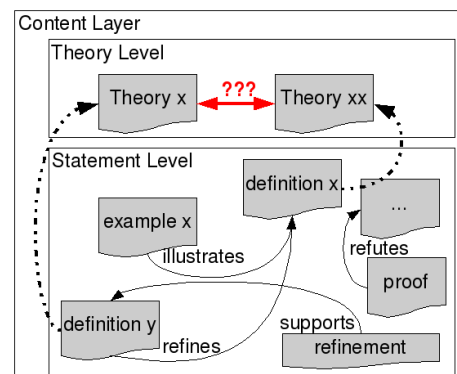


Figure 2: Reification of Theoretical Assumptions

Reifying the Choice of Examples

In order to provide the *reification of choice*, we draw on the *notion of variants*, i.e. *alternative choices* on the content layer of documents, in particular the statement layer. We draw on our previous work on the *representation of variants* in OMDOC [KMM07a]. For example, users can be facilitated to create rather *abstract documents*, which include alternative examples or illustration. These *variant documents* can be *adapted* or *substantiated* depending on specific *variants specifications* such as the language (German, English), the area of application (mathematics, physics, computer science), or the level of detail (short or long version). In Figure 3, the *example 1* node points to four possible example on the content layer. These are annotated with their *variant values*: For example, *example x* is in *English* and a *short variant*; while *Beispiel xx* is in *German* and a *long variant*. All variant examples are inter-related via a *variant dimensions* such as *translation* or *level of detail*. The selection of a concrete example depends on the given *variant specification*, e.g. *English* and *short*. In our case, *example x* is selected.

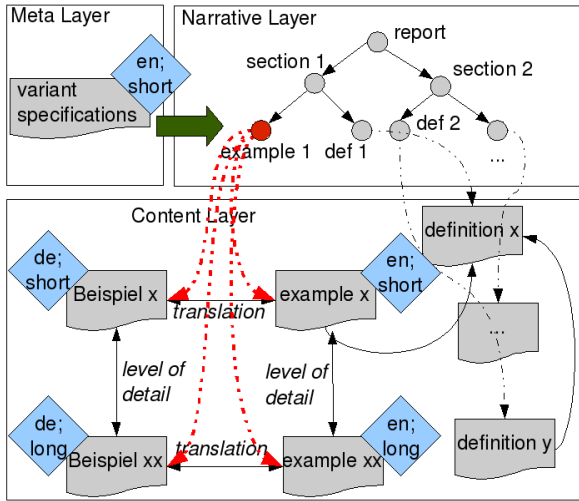


Figure 3: Reification of Choice

Reifying Notation Practice

In [KMR08; KMM07b] we reified *notation preferences* of scientists into *notation specifications* [KMR08], which are applied to the meaning of artifacts, i.e. the content layer, to generate user-specific presentation of documents. For example in Figure 4, two different notation specifications are applied to the *definition* of the *binomial coefficient*. The resulting fragments differ in the notation for binomial coefficient, i.e. C_k^n versus $\binom{n}{k}$, the highlighting of symbols and the layout of variables (cf. Section 6.1).

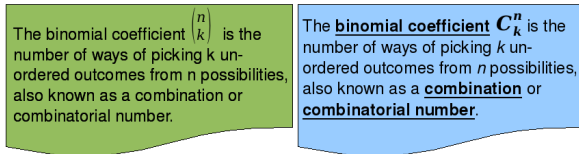


Figure 4: Reification of Notation Practice

Reifying Structural Preferences

We can reify the *structure (or outline)* of artifacts based on our *narrative-content model* [KMM07b] or, alternatively, by integrating other *outline markup language* such as [OPM08] (cf. Section 6.2).

5.3 Granularity

The previously mentioned approaches in Section 3 are limited to computations on document-layer. In contrast, semantic markup facilitates a more granular computation of document fragments and, thus, opens up new spaces for knowledge management on the web.

We make use of our *module system MMT* [RK; Rab08] to provide a web-scalable naming scheme: Documents and their fragments (on all layers and levels) do not have to be stored in the same memory or on the same machine, but are accessibly via globally unique URIs. This allows for a granular addressing and referencing of documents and document fragments stored in MMT-aware databases of the WWW.

The granular naming scheme also facilitates the granular markup of relations as well as annotations of fragments. Consequently, we elaborate approaches, which are limited to relations or metadata on document-level and can refine existing methods, such as citation analysis, to compute relations between document fragments on arbitrary granular levels. For example in Figure 5, the granular references between the fragments inside the technical report and slides as well as across both documents are illustrated.

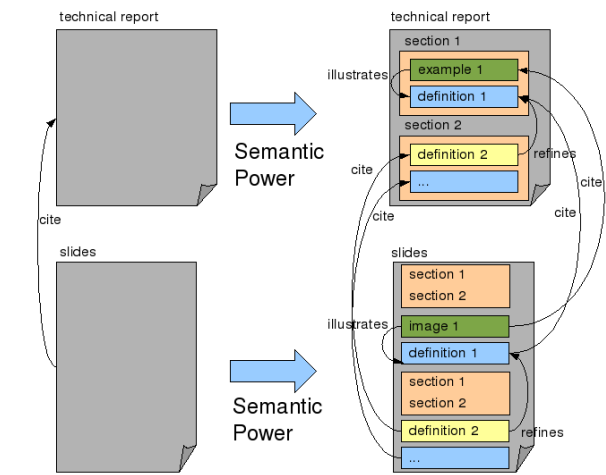


Figure 5: Granular Referencing and Addressing

5.4 Novel Services

The semi-formal markup allows mathematical software systems to *provide novel services*: For example, [Rab08] provides the *translation of logics*, [Lan08] and [pan08] build on the ontological relations to support *collaborative editing, review, and discussion* of documents, [loc07; MK08b] makes use of semantics to provide *granular change management*, and ACTIVEMATH provides *semantic learner models*. Our vision is to *interpret* collections of semantically marked up artifacts to model *virtual CoPs*, their *common repertoire*, and, particularly, their *common preferences*, which eventually facilitate the selection, structuring, and adaptation of information.

6 Virtual CoPs for CoPing

We propose the interpretation of semantically marked up artifacts to compute the *differences* between user (or rather between the user's artifacts and their interrelations) to eventually build *parametrized clusters* of similar users, henceforth referred to as *virtual CoPs*. The parameters define different *dimensions* for the clustering, e.g. the common *basic assumptions* or *background*, the common *choice of examples*, or the common *notation preferences*. We do not claim that the computed clusters (or virtual CoPs), are CoPs wrt. to Lave and Wenger [LW91] as they only consider selected dimensions. However, they provide initial means for other users to cope with information without prior interactions.

6.1 Semantic Differencing and virtual CoPs parametrized by Notation Practice

In the following sections, we compute the similarity of notation preferences of two users and compute their virtual CoP's common notation practice.

Listing 1: The Representation for $\binom{n}{k}$

```

<notation name="binomial">
  <prototype>
    <om:OMA>
      <om:OMS cd="combinat1" name="binomial"/>
      <expr name="arg1"/>
      <expr name="arg2"/>
    </om:OMA>
  </prototype>
  <rendering context="language:German">
    <m:mfrac linethickness="0">
      <render name="arg1"/>
      <render name="arg2"/>
    </m:mfrac>
  </rendering>
</notation>



```

Representation of Notation preferences

Notation preferences are tuples of the meaning of a notation (denoted by the `prototype`) and the actual presentation (denoted by the `rendering`). For example, the notation of a binomial coefficient is a tuple of the meaning, i.e. the number of ways of picking k unordered outcomes from n possibilities, and a potential presentation, e.g. $\binom{n}{k}$.

Listing 1 provides the XML-presentation of the notation specification of the binomial coefficient: The `prototype` element includes the content-representation of the notation in OPENMATH⁴, while the `rendering` element provides the *mapping* from the OPENMATH representation to Presentation-MATHML [W3C03]. Moreover, the `rendering` element has an optional `context`-attribute to specify the *context* of the notation.

The Figure to the right illustrates the notation preferences of two users A and B . A 's notation specification includes a notation for the *binomial coefficient* $\binom{n}{k}$, a notation for *multiplication* $*$, and a notation for the *cross product* \times . B 's notation specification comprise a different notation for the *binomial coefficient* C_k^n , two notations for *multiplication* \times and $*$, no notation for addition and cross product, but a notation for *subtraction* $-$ and the *cartesian product* \times . In the following section, we compute the similarity of A 's and B 's notation preference.

Notation pref. of A  and B 

| | |
|----------------------|----------------------|
| Binomial coefficient | Binomial coefficient |
| $\binom{n}{k}$ | C_k^n |
| multiplication | multiplication |
| * | \times * |
| Addition | Subtraction |
| + | - |
| Cross Product | Cartesian Product |
| \times | \times |
| prototype | rendering |

Semantic Difference of Notation Preferences

To compute the difference between notation preferences, we make use of our *model-based diff, patch, and merge system (mdpms)* [mdp08]: The system takes as input two notation specifications in XML as well as *user-defined equality relations* and returns a *semantic difference (semantic diff)* [LM08]. The equality relations of `mdpms` allows to parametrize the differencing, i.e. permits the definition of equality based on different dimensions: For example, two notation specifications are equal, if their concepts (prototypes) and notations (rendering) are equal, ignoring the context annotations. Moreover, the equality relation may ignore or define an order of elements, as e.g. notation specifications are unordered lists of `notation` elements.

Listing 2 provides the *semantic diff* of the notation specification from A to B , which we *interpret* to describe the differences in A 's and B 's notation practice: For example, a *remove-statement* can indicate that A *knows* a concept or notation that B is not aware of. Vice versa, and *append-statement* or *insert-statement* signifies that A is still missing certain background in order to understand B 's notation system. Moreover, difference in concepts (prototypes) are more *severe* than difference in notations (renderings). The former means that a users is missing the meaning and understanding of a mathematical object, while the latter only states that he is used to different notations but might be able to transfer his previous experiences to adjust to other notations.

⁴Prototypes can also include Content-MATHML [W3C03] representations.

Listing 2: *Semantic Diff* based on notation practice

```
<?xml version="1.0"?>
<xupdate:modifications version="1.0"
  xmlns:xupdate="http://www.xmldb.org/xupdate">
  <!--remove notation of A for binomial coefficient -->
  <xupdate:remove select="/ notations / notation [1]/ rendering" />
  <!--add notation from B for binomial coefficient -->
  <xupdate:insert -after select="/ notations / notation [1]/ prototype" >
    <xupdate:element name="rendering">
      ...
    </xupdate:element>
  </xupdate:insert -after>
  <!--add notation from B for multiplication -->
  <xupdate:insert -before select="/ notations / notation [2]/ rendering" >
    <xupdate:element name="rendering">
      ...
    </xupdate:element>
  </xupdate:insert -before>
  <!--remove addition -->
  <xupdate:remove select="/ notations / notation [2]" />
  <!--remove cross product -->
  <xupdate:remove select="/ notations / notation [3]" />
  <!--append subtraction and cartesian product -->
  <xupdate:append select="/ notations " child="last()">
    <xupdate:element name="notation">
      <xupdate:attribute name="name">subtraction</xupdate:attribute>
    </xupdate:element>
    <xupdate:element name="notation">
      <xupdate:attribute name="name">crossProd</xupdate:attribute>
    </xupdate:element>
  </xupdate:append>
</xupdate:modifications >
```

However, missing entries in the notation specification of A do not necessarily mean that A doesn't know a concept or notations, but could also be interpreted as different *interest* or *focus*, i.e. A simply *doesn't like* or *uses* a specific notation but very well knows about it. The interpretation of *semantic diff* therefore depends on the users as well as his current context and needs careful considerations.

In addition to our interpretation of the *semantic diff*, we also need a *mapping* to a *numeric similarity measure* to use our computed difference in a *standardized clustering algorithm* [BSMG02] and to eventually compute clusters of similar users, which we want to call *virtual CoPs*. For illustration we use the formulae below to compute a similarity measure d , which considers the *number of actions*⁵ as well as their *weighted type* and *depths* in the *semantic diff*.

$$d = \sum_{k=1}^n \frac{w(\text{type}(a_k))}{\text{depth}(a_k)}$$

A value of $d = 0$ signifies the full equality of two notation specifications. By dividing an action's value with its depth, we weight modification of *deeper* elements in the *notation specification tree*, e.g. the inserting of additional notations for a known concept, lower than modification on higher level e.g. the appending of an unknown concepts. For our example, we compute a numeric difference based on $n = 6$ actions:

$$d = \frac{w(\text{rm})}{3} + \frac{w(i)}{3} + \frac{w(i)}{3} + \frac{w(\text{rm})}{2} + \frac{w(\text{rm})}{2} + \frac{w(\text{app})}{1}$$

Given that all actions are *equally weighted* with $w(a_k) = 1$, we result in a measure $d = 3$. However, please note that the previous computation is for illustrative purpose only and needs to be evaluated.

Common Notation Preferences of CoP's

In order to compute the common preference of A and B we make use of a *semantic similarity*: For this we compute the *similarity* between two XML-representation based on

⁵Actions are all activities in the semantic diff from A to B that are required to update A towards B . In our example, these include remove, insert-after, insert-before, and append statements.

a given *equality relations*. For our example the common notation practice comprises one specifications, since A and B only share the understanding and notation of the *multiplication* concept. However, the equality relations for the *semantic similarity* could be based on equal prototypes, ignoring different renderings and context annotations. In this case, we simply consider all equal prototypes and merge the respective renderings. Consequently, the common notation specification would include the concepts *binomial coefficient* and *multiplication* with two notations each.

6.2 Semantic Differencing of Structures

In the following section we illustrate our (revised) narrative-content model [KMM07b] and the semantic differencing of narrative structure. We then point to other structures that are suited for the description of a CoP's shared interest or expertise.

Representation of Narrative Structures

Listing 3 provides the representation of the *lecture notes* on General Computer Science of user A . The narrative elements are used to structure the course material into sections and subsections. They include `metadata` as well as `ref` elements, which point to course content on the web.

Listing 3: Narrative Structure of the GenCS Lecture

```
<omdoc xml:id="gens07" type="lecture">
  <metadata><dc:title>General Computer Science</dc:title></metadata>
  <narrative type="section">
    <metadata><dc:title>Discrete Math</dc:title></metadata>
    <narrative type="subsection">
      <metadata><dc:title>Natural Numbers</dc:title></metadata>
      <ref xref="http://panta-rhei.kwarc.info/slides/natnums-intro"/>
    </narrative>
    <narrative type="subsection">
      <metadata><dc:title>Naive Sets</dc:title></metadata>
      <ref xref="http://panta-rhei.kwarc.info/slides/native-sets"/>
    </narrative>
  </narrative>
  <narrative type="section">
    <metadata><dc:title>Computing Functions</dc:title></metadata>
    <narrative type="subsection">
      <metadata><dc:title>Datatypes</dc:title></metadata>
      <ref xref="http://panta-rhei.kwarc.info/slides/datatypes"/>
    </narrative>
    <narrative type="subsection">
      <metadata><dc:title>Abstract Datatypes</dc:title></metadata>
      <ref xref="http://panta-rhei.kwarc.info/slides/adt"/>
    </narrative>
  </narrative>
</omdoc>
```

Semantic Difference of Narrative Structures

The user B restructures the material for her course on General Computer Science as follows: (1) Computing Functions, (1.2) Abstract Datatypes, (1.3) Datatypes, (2) Discrete Math, (2.1) Natural Numbers, and (2.2) Naive Sets.

Listing 4: A *Semantic Diff* of Course Structures

```
<?xml version="1.0"?>
<xupdate:modifications version="1.0"
  xmlns:xupdate="http://www.xmldb.org/xupdate">
  <!--remove subsection "abstract datatypes" from B -->
  <xupdate:remove select="/omdoc/narrative [1]/ narrative [1]" />
  <!--insert subsection "abstract datatypes" after "datatypes" -->
  <xupdate:insert -after select="/omdoc/narrative [1]/ narrative [1]" >
    <xupdate:element name="narrative">
      ...
    </xupdate:element>
  </xupdate:insert -after>
  <!--remove section "discrete math" from B -->
  <xupdate:remove select="/omdoc/narrative [2]" />
  <!--insert section "discrete math" of A -->
  <xupdate:insert -before select="/omdoc/narrative [1]" />
    <xupdate:element name="narrative">
      ...
    </xupdate:element>
  </xupdate:insert -before>
</xupdate:modifications >
```

We apply `mdpm` to compute the *semantic diff* from A 's to B 's structure (cf. Listing 4), which is mapped to a numeric similarity measure. The *equality relation* ignores metadata elements, but considers the order of narrative elements.

We consider $n = 4$ actions to reorder B 's structure towards A 's outline. Given that all actions are equally weighted we result in the following numeric measure:

$$d = \frac{w(rm)}{3} + \frac{w(i)}{3} + \frac{w(rm)}{2} + \frac{w(i)}{2} = 1.67$$

Common Structures of virtual CoPs

Instead of using metadata of an author's documents, we can extract the concepts he is using in order to model his expertise. Our expertise models include pointers [Mel01; KBB08] to the OPENMATH content dictionaries [OMC] and the concepts extracted from a user's documents, such as his notation specification file. For example, a pointer to `http://www.openmath.org/cd/combinat1.xhtml#binomial` represents a user's familiarity with the concept *binomial coefficient*. Since expertise model of A and B are simple lists of pointers and, thus, trivial structures, the computation of differences is rather a counting of pointers that are not shared by A and B ; while their virtual CoP would be describe by the intersection of their pointer-lists. However, the discussion of expertise is more exciting if we think back to our example in Figure 2. A *similarity measure* based on the matching of *theory pointers* identified A and B as unequal wrt. to their familiarity of the two theories *theory x* and *theory xx*. However, with our semantic markup we are able to relate theories. Given that a *mapping* from *theory x* to *theory xx* exists, A and B would be equal with respect to their *common theoretical assumptions*. Although our illustrations is based on expertise model, our approach can easily be applied to other (simple) structures such as bookmarks or news.

7 Conclusion & Outlook

We have illustrated how semantic technologies facilitate the reification and extraction of scientific practice. Moreover, we discussed how our approach facilitates the propagation of similarities on content, structure, and presentation layer to the social layer: For example, similar notation preferences or the use of the same mathematical concepts allow to identify similarities among users and the clustering of virtual CoPs.

Vice versa a user's or CoP's representation can define the content, structure, and presentation of artifacts: For example, a notation specification of A can be applied to a document of B , allowing A to more easily understand B 's illustrations. Alternatively, expertise and narrative representation can be used to select or structure a user's artifacts. Moreover, the common preferences of CoPs allow new users to select, structure, and adapt presentation without prior investments or initialization of their user models.

For the future, we want to substantiate our approach by extending and testing our *visionary ideas* in concrete use cases (cf. [Mül08]). We also want to evaluate whether semantics elaborates existing approach based on informal documents. For example, most citation management approaches are limited to only one type of document interrelations, i.e. the *cite* relation, on document-level. In contrast, semantic markup explicates several types of relations and document fragments on granular level.

Acknowledgement

This work was funded by the JEM-Thematic-Network ECP-038208. The author would further like to thank the members of the KWARC group and the DFKI Bremen for our fruitful discussions and their continuous feedback.

References

- [Act07] ACTIVE MATH, seen February 2007. web page at <http://www.activemath.org/>.
- [AT05] Gediminas Adomavicius and Alexander Tuzhilin. Toward The Next Generation Of Recommender Systems: A Survey Of The State-Of-The-Art And Possible Extensions. *IEEE Transactions on Knowledge and Data Engineering*, 17(6):734–749, 2005.
- [BSMG02] David J. Bartholomew, Fiona Steele, Irini Moustaki, and Jane I. Galbraith. *The analysis and interpretation of multivariate data for social scientists*. Chapman & Hall/CRC, 2002.
- [CC08] Jean-Philippe Cointet and David Chavalarias. Multi-level science mapping with asymmetrical paradigmatic proximity. In *Networks and Heterogeneous Media (NHM)*, volume 3, pages 267–276. <http://aimSciences.org>, 2008.
- [HG07] Brent Hendricks and Adan Galvan. The Connexions Markup Language (CNXML). <http://cnx.org/aboutus/technology/cnxml/>, 2007. Seen June 2007.
- [KBB08] Christopher M. Kelty, C. Sidney Burrus, and Richard G. Baraniuk. Peer review anew: Three principles and a case study in postpublication quality assurance. *Proceedings of the IEEE; Special Issue on Educational Technology*, 96(6):1000–1011, June 2008.
- [KK08] Andrea Kohlhasse and Michael Kohlhasse. Semantic Knowledge Management for Education. *Proceedings of the IEEE; Special Issue on Educational Technology*, 96(6):970–989, June 2008.
- [KM07] Rushed Kanawati and Maria Malek. Computing Social Networks for Information Sharing: A Case-Based Approach. [LNC07], pages 86–95.
- [KMM07a] Michael Kohlhasse, Achim Mahnke, and Christine Müller. Managing Variants in Document Content and Narrative Structures. pages 324–229, 2007.
- [KMM07b] Michael Kohlhasse, Christine Müller, and Normen Müller. Documents with flexible notation contexts as interfaces to mathematical knowledge. In Paul Librecht, editor, *Mathematical User Interfaces Workshop 2007*, 2007.
- [KMR08] Michael Kohlhasse, Christine Müller, and Florian Rabe. Notations for living mathematical documents. In *Mathematical Knowledge Management, MKM'08*, LNAI. Springer Verlag, 2008. in press.
- [Koh05] Michael Kohlhasse. Semantic markup for $\text{T}_{\text{E}}\text{X}/\text{L}_{\text{A}}\text{T}_{\text{E}}\text{X}$. Manuscript, available at <http://kwarc.info/software/stex>, 2005.
- [Koh06] Michael Kohlhasse. OMDOC – An open markup format for mathematical documents [Version 1.2]. Number 4180 in LNAI. Springer Verlag, 2006.
- [Kuh96] Thomas S. Kuhn. *The Structure of Scientific Revolution*. University Of Chicago Press; 3 edition, 1996.
- [KW05] Andrea Kienle and Martin Wessner. Principles for Cultivating Scientific Communities of Practice. In Peter van den Besselaar, Giorgio de Michelis, Jenny Preece, and Carla Simone, editors, *Communities and Technologies*, pages 283–299. Springer Verlag, 2005.
- [Lan08] Christoph Lange. SWiM: A semantic wiki for mathematical knowledge management. web page at <http://kwarc.info/projects/swim/>, seen February 2008.
- [LM08] Andreas Laux and Lars Martin. XUpdate: Xml update language, seen May 2008. XML:DB working draft at <http://www.xmldb.org/xupdate/xupdate-wd.html>.
- [LNC07] Lecture Notes in Computer Science: Online Communities and Social Computing, 2007.
- [loc07] *locutor*: An Ontology-Driven Management of Change, seen June 2007. system homepage at <http://www.kwarc.info/projects/locutor/>.
- [LW91] Jean Lave and Etienne Wenger. *Situated Learning: Legitimate Peripheral Participation (Learning in Doing: Social, Cognitive and Computational Perspectives S.)*. Cambridge University Press, 1991.
- [LWA08] *Wissens- und Erfahrungsmanagement LWA (Lernen, Wissensentdeckung und Adaptivität) Conference Proceedings*, 2008.
- [mdp08] *mdp*: A Collection of Model-based DIFF, PATCH, MERGE Algorithms, seen March 2008. system homepage at <http://www.kwarc.info/projects/mdp/>.
- [Mel01] Erica Melis. User model description. DFKI Report, DFKI, 2001.
- [MK08a] Christine Müller and Michael Kohlhasse. Towards A Community of Practice Toolkit. In Christine Müller, editor, *Proceedings of the 2nd SCooP Workshop*, 2008.
- [MK08b] Normen Müller and Michael Kohlhasse. Fine-Granular Version Control & Redundancy Resolution. [LWA08]. submitted.
- [Mül06] Normen Müller. An Ontology-Driven Management of Change. In *Wissens- und Erfahrungsmanagement LWA (Lernen, Wissensentdeckung und Adaptivität) conference proceedings*, pages 186–193, 2006.
- [Mül08] Christine Müller. Towards the Adaptation of Scientific Course Material powered by Community of Practice. [LWA08]. in submission.
- [OMC] OPENMATH content dictionaries. web page at <http://www.openmath.org/cd/>.
- [Ope07] OPENMATH Home. Web site at <http://www.openmath.org/>, seen March 2007.
- [OPM08] Outline Processor Markup Language. <http://www.opml.org/>, seen June 2008.
- [pan08] The panta rhei Project. <http://kwarc.info/projects/panta-rhei/>, 2008.
- [Rab08] F. Rabe. *Representing Logics and Logic Translations*. PhD thesis, Jacobs University Bremen, 2008. To appear.
- [RK] Florian Rabe and Michael Kohlhasse. A web-scalable module system for mathematical theories. Manuscript, to be submitted to the Journal of Symbolic Computation.
- [W3C03] W3C. Mathematical Markup Language (MathML) Version 2.0 (Second Edition). <http://www.w3.org/TR/MathML2/>, 2003. Seen July 2007.
- [Wen05] Etienne Wenger. Communities of Practice in 21st-century organization, 2005.
- [WM07] Marc Wagner and Christine Müller. Towards Community of Practice Support for Interactive Mathematical Authoring. In Christine Müller, editor, *Proceedings of the 1st SCooP Workshop*, 2007.
- [ZYAQ07] Jun Zhang, Yang Ye, Mark S. Ackerman, and Yan Qu. SISN: A Toolkit for Augmenting Expertise Sharing Via Social Networks. [LNC07], pages 491–500.