LECTORA

Towards An Interactive And Collaborative Reader For Mathematical Documents. *

Christine Müller

Jacobs University Bremen
D-28717, Bremen, Germany
c.mueller@iu-bremen.de

The explosive growth of information available online and the rapid increase of scientific information critically degraded the ability of scientists to keep informed about areas of interest other than their major research areas. The problem of information overload has been addressed by several approaches, such as providing personalized information by recommender systems, improving search by facetted browsing environments, and facilitating the assessment of information by means of annotations. Nevertheless, none of the existing methods provides a complete and/or satisfying solution to the information overload of scientists, in particular mathematicians. This work therefore aims at diminishing the overload of users, which are interested in mathematical information. It is further assumed that applying the economic theory of Communities of Practice (CoP) by Jean Lave and Etienne Wenger [LW91] to the domain of mathematics constitutes a promising approach to address this challenge. Hence, groups of users, e.g. researchers and students, with similar tasks and practices are modeled to make predictions on the relevance as well as the user's preferred content, structure, and appearance of a mathematical document. The user's practices are inferred from his ratings, annotations, and browsing behavior. Furthermore, it is assumed that some of his practices are inscribed into the user's collection of documents and that semantic representation formats facilitate their automated identification. Last but not least, an interactive, collaborative reader for mathematical documents, henceforth referred to as LECTORA, is implemented to verify this approach.

*The author would like to thank Michael Kohlhase, Andrea Kohlhase, Immanuel Normann, Normen Müller, and Christoph Lange for stimulating discussions in the early stages of this work.

Contents

1 Introduction

The explosive growth of information available online and the rapid increase of scientific information critically degraded the ability of researchers to keep informed: For example, in order to trace the discourse of fields of interest other than their major research area, scientist have to face a high rate of new material. Furthermore, they are forced to dig through a mass of existent information when looking at new areas. The problem of information overload has been addressed by the eLearning community by asking for a new ability by the user: *media competence*, i.e. an individual's ability to critically reflect information (cf. [GG99]). One way to acquire media competence is by means of other people's assessment of information in terms of its relevance, correctness, quality, usefulness, or trustworthiness.

Recommender systems support the challenge of information assessments, e.g. by means of personalized services for improved information access. One of the best known systems for recommendation based on user behaviors and ranking is the commercial online bookshop Amazon.com [LSY03]. The approach used by Amazon.com recommends similar shopping items based on the user's current purchase and the shopping behavior of other customers, i.e. by identifying which items they have purchased together. The approach requires minimal additional input of users, but requires a critical mass of purchases, so that new or unusual items are recommended less often. In contrast, [BP00] proposes the intelligent information agent DailyLearner for automatically personalizing a user's access to news, i.e. information with a high frequency. The agent considers the user's explicit ratings of text documents as well as implicit information based on their browsing behaviors. In addition, [BP00] suggests the analysis of collaborative annotation, i.e. the style, timeliness, and usefulness of news, in order to improve the quality of recommendations. The research on collaborative user ratings and annotations, has raised the interest of other commercial services, such as the online DVD rental service Netflix [Net07], which launched an open contest for improving its prediction of the users' movie preferences. The current approach clusters similar movies and analyzes how customers have rated them. Users who have given similar ratings to the same movies in a cluster are matched to like-minded viewers, i.e. as members of a specific community of interest. Based on these clusters customers have rented from in the past, other titles of the cluster are determined that customers have not yet rented. Subsequently, the system filters and recommends only those movies that have been highly rated by matched viewers. To sum up, the pre-introduced recommender systems all aim at identifying similarities, i.e. shared practices, among users, i.e. similar shopping and browsing behaviors as well as rankings, to describe communities of users, which the systems then base their recommendations on. Another approach for managing the general information overload of users is the personalized information presenter mSpace [msSO$^+$05], which applies facetted browsing technologies to improve the ordering and search for information based on each user's associations and browsing behavior. Alternatively, the reference management service Connotea [con07] focuses on the improvement of scientific information access. The system applies social computing methods by gathering individual annotations of users and by publishing these annotation to allow other users to assess existent and new scientific information more efficiently.

User practices, i.e. browsing behaviors, ratings, and annotations, also differ among scientific communities. For example, mathematical communities can be distinguished by their assessment of relevance, correctness, and originality of mathematical publications. However, this work assumes that numerous mathematical practices, i.e. choosing a specific notation, are inscribed into scientific documents (cf. [KK06]). Thus, user practices may not only be gathered by direct user feedback, i.e. ratings and associations, or by tracing the user, i.e. his browsing behavior, but may also be detected by analyzing his document collection, i.e. the set of documents a user writes or reads. Therefore, the more structured the respective format of the documents, the more practices can be identified. Hence, this work uses a semantic representation format for mathematical documents, i.e. Open Mathematical Documents OMDoc [Koh06], which facilitates the automated identification of practices and, thus, reduces the user's investment. Then the information on users is used to identify similarities, i.e. shared practices, that allow to describe specific scientific communities. In order to model practices and communities, this work proposes to apply the economic theory of Communities of Practice (CoP) by Jean Lave Etienne Wenger [LW91] to improve access to mathematical information based on prediction, i.e. community-aware recommendations, search rankings, and the adaptive presentation of mathematical documents.

This thesis consists of two phases. The first phase is a preparation phase in which practices of researchers, scholars, and students are formalized and user-specific practices are modeled. At the end of the phase, the first version of the reader, i.e. LECTORA LIGHT is implemented and used within an educational case study to realize and evaluate the previously made assumption whether practice are inscribed into documents and semantic representation formats may facilitate the automatic detection of practices (cf. section 4.1). The second phase is focuses on the modeling of communities, the application of CoPs to provide community-aware service as well as a case study and the implementation of an improved version of the reader, called LECTORA (cf. section 4.2).

The structure of this proposal is as followed: Section 1.1 exemplifies the introductory discussion by means of an example. Afterwards the approaches of this thesis are presented in section 1.2, followed by the enumeration of the research objectives in section 1.3. Section 2 includes a discussion of the related work. The theory of CoPs as well as mathematical practices are discussed in section 3.2. Finally, the three phases are illustrated in section 4.

1.1 Running Example

To exemplify the introductory discussion, imagine the following scenario: Immanuel is a mathematical scholar with a specific research interests and practices. For example, he is used to certain mathematical notations, bases his proof on various assumptions, and has an individual view on what documents are relevant for his research area and which mathematicians may relate to his work. Part of his practices are inscribed into the documents he writes, i.e. the material for a lecture. In addition, he continuously has to cope with new practices inscribed in the documents he reads, e.g. notations used in the field of electrical engineering that differ to the ones in mathematics. In his specialized area, Immanuel can draw on a wide background, numerous experiences, and various researchers in his personal network. For example, Immanuel can address his colleague Florian in the field of logic. However, he may not be able to draw on previous experiences or contacts in order to learn about the economic concept of communities of practice, although it might be very relevant for his current research on improving mathematical information access. In consequence, Immanuel may face difficulties to keep up-to-date with the research on CoPs. Thus, he has to invest a significant effort for searching, reading, and understanding respective publications in order to be able to judge upon their relevance and to adopt practices. In contrast, Normen, a first year undergraduate student of mathematics, wants to recapitulate Immanuel's lecture but has difficulties in understanding some of the material, especially his notation practice. Moreover, Normen tries to find additional examples and more detailed explanations of Immanuel's proofs.

To sum up, the following two problems arise:

1. How can Immanuel be supported to trace the discourse of communities other than his own field of specialization?
2. How can Normen be supported to most easily access/ learn about basic mathematical information?

1.2 Proposed Approach in this Thesis

This work focuses on two main goals: (1) Building an interactive and collaborative reader for researchers, scholars, and students that supports a more efficient access of mathematical documents. (2) Using the prototype to test the assumptions that (2.a) mathematical practices are inscribed into documents and that a semantically structured representation format supports the automatic detection of practices, and that (2.b) the descriptive theory of communities of practice by Lave and Wenger [LW91] can be applied to model mathematical communities which can then be used to provide community-aware services such as recommendations, search ranking of mathematical search results, and the adaptive presentation of mathematical documents.

Before going into details, the main methods of this work shall be discussed:

Communities of Practice This work proposes to apply the concept of Communities of Practice (COP) by Jean Lave and Etienne Wenger [LW91], a well-known and widely accepted theory in economics, to the field of mathematics. By definition, communities are groups of people that collaboratively develop knowledge, share experience, and build up a common identity. One occurrence of communities are online communities which are groups of people that share similar interests and exchange experiences via websites, e.g. Wikipedia [Wik06]. In contrast, communities of practice are informal practice-oriented groups of people with similar tasks. They are further described by a specific set of practices, which are shared by its members to achieve these tasks. A typical example for a community of practices are the members of the KWARC community[1] [KWA07], which share the task of researching on knowledge representation and the development of technologies for semantic markup of technical documents. In consequence, the members of the KWARC community share similar practices, such as the use of a semantic representation format, i.e. the Open Mathematical Document Format (OMDOC) [Koh06], the open communication structure, a flat hierarchy, and weekly discussions and meetings.

The concept of COPs has been widely used to describe the process of social learning that occurs when people who have a common interest in some subject or problem collaborate over an extended period to share ideas, find solutions, and build innovations. Several existing COPs have been analyzed, but so far no attempts of modeling COPs and using them to predict memberships, interest, and practices of its members and, based on that, to provide community-aware services, have been made. This work applies the descriptive theory of COPs to implement a community-aware reader prototype for mathematical documents and aims at using COPs to make predictions on the relevance of documents as well as the user's preferred content, structure, and appearance of a mathematical document.

The domain of mathematics was chosen since we believe it to be the language of science, i.e. mathematics is not only a domain of science, but the basis for other science, e.g. natural science, engineering, and computer science (cf. [JRS05]). Thus, by basing this work on mathematics, we hope to achieve benefits for other research domains such as physics or chemistry. Keeping that in mind, this thesis eventually contributes to the research on extending the Open Mathematical Document Format OMDOC to general science. In the respective research projects (SCIML and DOCKON), mathematics is defined to be the meta language of science that may be easily used to specify other science mark-up languages such as PHYSML [HKS06], an OMDOC extension for physics. Furthermore, mathematics is assumed to be directed towards a very wide and heterogenous auditorium with various interests and practices. In consequence, various mathematical communities and especially their heterogenous practices, i.e. the professional way of doing mathematics, can be identified and eventually be used to gain findings on the application of COPs to other sciences and less formal research domains. In order to visualize mathematical practices, the work of Heintz [Hei00] and Polya [Pol73]) is discussed and conclusions for the identification and modeling of user expectations and characteristics of communities are drawn (cf. section 3.2).

The Design Method: An Added-Value Analysis Andrea Kohlhase and Normen Müller [KM07] propose a new design method, the Added-Value Analysis, for identifying desirable services of semantic work environments (SWE). Kohlhase and Müller suggest that the identification of services that evolve in the process in which individual users use a system, increases the user's motivation to invest essential

[1] KWARC is the 'Knowledge Adaptation and Reasoning for Content' research group at the Jacobs University Bremen and the author's work group, i.e. one of her communities of practice

efforts and taking respective actions within the resulting environment. The method of an Added-Value Analysis is applied in the case studies to identify desirable features for users of the interactive reader LECTORA LIGHT and the improved prototype version LECTORA (cf. section 5).

Prototypes and Case Studies The thesis will be divided into three main phases:

- *Phase I: Practice-Oriented Phase.* The first phase is a preparation phase in which practices of researchers, scholars, and students are formalized and user-specific practices are modeled. The first phase aims at building a light version of the reader that uses the information on individual user practices to provide practice-oriented services. The phase is completed by a case study based on a students and educational content of a Computer Science lecture and the implementation of a prototype (LECTORA LIGHT).
- *Phase II: Community-Aware Phase.* The second phase is the most interesting phase from the research perspective since it applies an descriptive economic theory to the domain of mathematics to predict the user's preferences. It focuses on the modeling of communities, the application of CoPs to provide community-aware service as well as the implementation of an improved version of the prototype (LECTORA).
- *Phase III: Evaluation.* The third phase is used to evaluate the findings and intuitions gained within phase I and II. These will be used to verify the two assumptions this work is based on: (1) Practices are inscribed into documents and a semantically structured representation format supports the automatic detection of practices and, thus, reduces the investment of users. (2) Applying the descriptive theory of Communities of Practice [LW91] improves access to mathematical documents by offering community-aware services. In addition, the Added-Value analysis by [KM07] is applied to contrast the user's benefits provided by LECTORA with the required additional investments of users, i.e. in form of an extended feedback and annotations.

1.3 Research Objectives

To sum up, the objectives of my PhD thesis are:

1. *Formalization of practices* of researchers, scholars, and students: Different practices shall be identified, structured, and represented (cf. section 4.1).
2. *Modeling of user* based on multi-dimensional criteria: It is assumed that part of these practices are inscribed into technical documents. These shall be detected automatically to reduce the user's investment. Other practices can be gained by tracing the user, i.e. by monitoring his browsing behavior, or alternatively by analyzing explicit feedback, i.e. annotations and ratings. By rule-based inference individual practices can be inferred from the formal representation of step one (cf. section 4.1).
3. *Providing practice-oriented services* that are restricted to the application of individual user practices rather than information on the user's CoPs (cf. section 4.1).
4. *Modeling communities* by finding shared mathematical practices, i.e. similarities among user models. Two methods, namely an intensional and extensional modeling approach for CoPs, will be discussed (cf. section 4.2).
5. *Providing community-aware services* to improve search ranking, recommendations, and presentation of mathematical documents (cf. section 4.2).

2 Related Work

The proposed approach of the interactive and collaborative reader LECTORA, which offers adapted recommendation, search rankings, and document adaptations, relates to several research areas that are discussed in this section.

2.1 Recommender Systems

Recommender systems provide personalized and context-oriented information, i.e. the system pre-selects information based on the user's profile or behavior and relates to the personalization of information access in LECTORA LIGHT. A survey by Gediminas Adomavicius and Alexander Tuzhilin [AT05] highlights various limitations of the current generation of recommender systems and discusses initial approaches for improving current implementations. They distinguish three categories for recommendation systems, namely content-based, collaborative, and hybrid approaches, all of which have various benefits and limitations (cf. [AT05]).

Content-Based Techniques work well on text documents, i.e. to recommend text items that are described by keywords, and are mostly measuring the frequency of specific keywords appearing in documents. In consequence, two different documents represented by distinguish between well-written and badly written documents. Further drawbacks of content-based recommendations are the required *investment by the user*, i.e. the need for users to previously rate a sufficient number of items before a recommender system can rank other items. Finally, systems can only recommend items that score high against a user's profile. Thus, items that are different to everything the user has seen before cannot be identified.

Collaborative Recommendations The approach driven by LECTORA relates to collaborative recommendations, e.g. basing recommendations on the rating of other users that are 'peers' of the user and rate items similarly. As opposed to content-based approaches, collaborative recommendations measure the similarity of user-specific ratings 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 problem of *new users* remain since the system first needs to learn the user's preferences to identify other related users. Furthermore, collaborative approaches require a critical mass of ratings to be successful. For example, items rated by only some users are recommended rarely or users with unusual preferences will not relate to many users, thus only receive poor recommendations. An attempt to solve these problems is *exploring similarities among users*, e.g. by identifying associations among users through their past transactions and feedback (cf. [AT05]).

Hybrid Approaches [AT05] also describe the *hybrid approach*, a combination of content-based and collaborative methods, and highlight the limitations of existing implementations. They suggest to support *multi-criteria ratings*, to implement an *improved understandings of users and items*, and to consider *contextual information* about the user, i.e. not to restrict recommendation to a two-dimensional space of user and item. Specifically, [AT05] propose including the point of time a recommendation is requested or a specific person the information is shared with, and to allow reducing recommendations to specific criteria. Furthermore, the *intrusiveness* of recommender systems is highly criticized: For example, instead of requiring explicit feedback and a significant level of involvement, systems should estimate ratings, e.g. by the time a user spends reading a document. However, estimation decreases the accuracy of recommendations and needs to be carefully implemented. [AT05] also stress the *inflexibility* of recommender systems that only allow predefined sets of recommendations and the need for measuring the effectiveness of recommendations in terms of *usefulness and quality* from the perspective of users rather than applying statistical methods to compute an objective accuracy and coverage of the system. Further issues emphasized are trustworthiness, explainability, and privacy issues.

To sum up, LECTORA LIGHT can benefit from existing recommendation approaches, e.g. the retrieval of documents based on collaborative user rankings and user models. However, she aims at augmenting existing recommender systems by formalizing and modeling various user practices. For example, the automatic detection of user practices based on the documents reduces the initial investment of users. In addition, LECTORA's approach takes the suggestions by [AT05] into account: For example, recommendation will be based on multiple criteria, i.e. sets of user practices including the usefulness and quality of documents. Furthermore, LECTORA LIGHT infers individual user practices from documents, behaviors, and annotations by applying rule-based inferences and a formal representation of practices. Hence, LECTORA can explain and justify its modeling processes.

2.2 Living Documents

An approach that relates to the personalized presentation of documents are *living documents*. Living documents are work environments that allow to dynamically change the content, structure, and appearance of documents. They furthermore provide facilities for annotating documents, such as indicating their quality, usefulness, relevance or correctness. The term *living document* refers to two aspects: (1) living layout and (2) living content of documents, both of which support the adaptive presentation of documents and shall be discussed in this section.

Living Layout The digitalization of documents has offered several ways of adapting the layout of a document. For example, LaTeX class and style files allow to transform a document's content in several ways. Alternatively, providing living layout has been addressed by the project LDoc [KMA06] in which lecture notes (or scripts) are extended by multimedia-based elements, such as videos, audio files, or graphics, to provide a flexible and innovative learning environment.

Living Content The rise of semantic technology has brought up several methods for representing machine-readable content and has provided services to automate e.g. the personalized retrieval of content as well as adapting its structure. For example, the semantic representation format Open Mathematical Documents (OMDOC) [Koh06] provides two abstraction layers, in particular Content OMDOC and Narrative OMDOC, that facilitate the adaptive selection and structuring of content. Figure 1 shows both layers as well as their interrelations: Content OMDOC represents mathematical content, which constitutes OMDOC content commons [CNX07] and their relations. For example, a definition depends on a mathematical theory. Two exercises are recommended in order to learn the definition. One of the exercises is linked to an example that may help solving it. In contrast, Narrative OMDOC specifies the structure of the content, i.e. the order of the content commons. For example, Normen wants to learn the definition and so LECTORA offers both exercises as well as the example. Conversely, Immanuel only needs to look at the definition and the theoretic assumption underneath in order to quickly pick up the mathematical idea behind it.

In addition, semantic representation formats, such as MATHML[Mat07], support the adaptation of mathematical notation by separately specifying the meaning (Content MATHML) and various alternative presentations (Presentation MATHML) of the notation. We can infer that, semantic technologies also augment the implementation of dynamic document layouts. Figure 2 shows the specification in MATHML that represents the mathematical meaning of a binomial coefficient as well as various ways to present it. Based on a semantic representation format, LECTORA adapts the notation to the preferences of a user, assures a consistent use of notation, as well as detects notation preferences based on the user's collection of documents.

ACTIVEMATH is a user-centered learning environment, which is based on the semantic representation format OMDOC: Thus, instead of only storing presentation-oriented content and providing static course

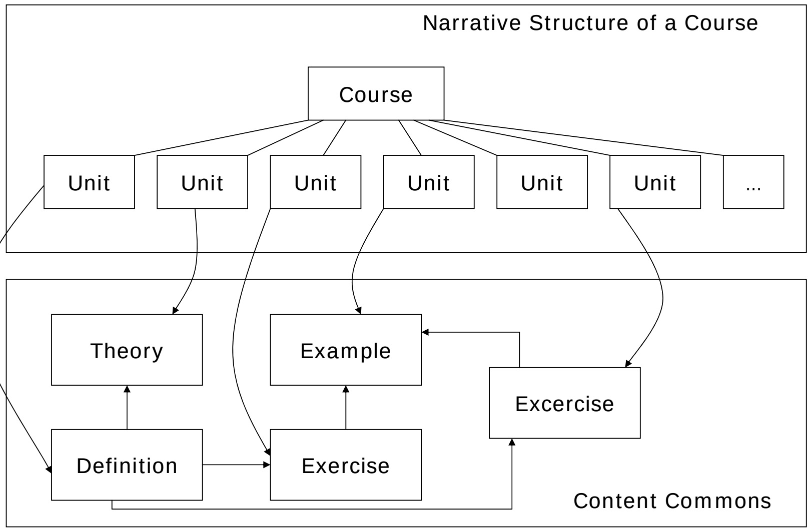

Figure 1: Content and Narrative OMDOC.

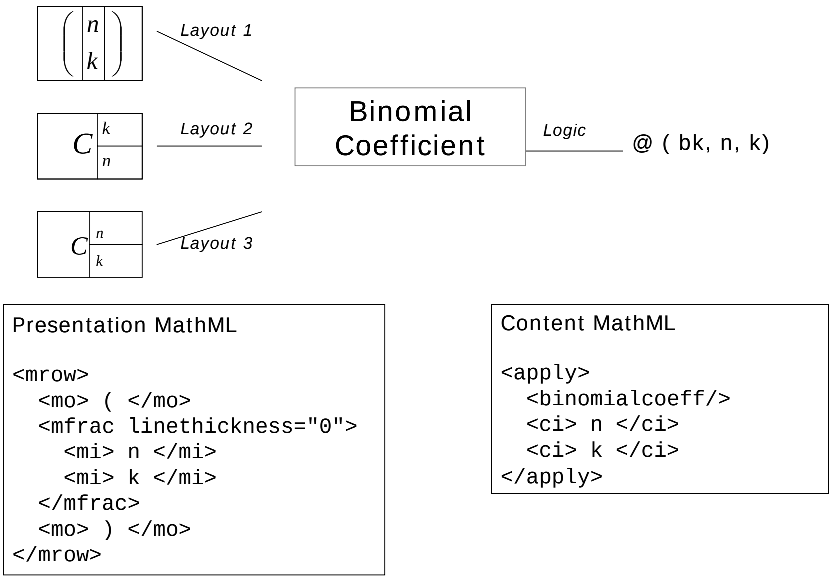

Figure 2: Meaning and presentation of a binomial coefficient.

structures, authors are required to supply semantically structured content, its presentation, and a narrative structure of the course as well as didactic requirements in form of pedagogical rules. The highly structured content is indispensable for the reusability of material and for the combination of material from different sources. The central component of ACTIVEMATH is a presentation planner, which generates personalized courses: Based on the user's previous interaction and profile, all relevant concepts and dependent information are collected from the OMDOC knowledge base. Afterwards, pedagogical knowledge is applied, i.e. heuristical rules are used to decide which information should be presented on a page and in which order the information should appear. Furthermore, these rules define the number of exercises and examples as well as their degree of difficulty (cf. [MAF$^+$01] and [LMU01]).

To model the user's profile and history, ACTIVEMATH collects static and dynamic information about students, such as their interaction history as well as their profile, i.e. the user's learning goals, field, scenario, knowledge mastery data, and appearance preference. Information on users is either manually specified or are inferred by the system by analyzing the user's interactions. During the registration, users are asked to describe their preferences, field of interest, their knowledge, comprehensibility, and application of concepts. Currently, users can only indicate their preferred appearance of a course in terms of color. However, the ACTIVEMATH group started to take attempts to find a presentation configuration that supports different groups of learners. Furthermore, the ACTIVEMATH research group currently works on the automatic translation of notations (cf. [MLUM05]). The field of a user, e.g. biology, mathematics, or statistics, has an impact on the type of examples and exercises the system includes in a course. The content of a course is influenced by the user's learning goal, i.e. a concept in the underlying mathematical knowledge base. Users can either enter their goal explicitly, i.e. by manually choosing a predefined course, or, alternatively, let the system generate a new course based on ACTIVEMATH's goal concept and the scenario a learner chooses: Several scenarios are possible, e.g. 'exam', 'exam preparation', 'overview', or 'guided tour' (cf. [LMU01]). The system's assumptions on the user's current *knowledge* are visualized by differently colored boxes of each course section. Furthermore, the user can access her *learning model*, i.e. mastery assessments of each exercise based on her performance. The assumptions about the mastery of an exercise is based on the correctness of its solution. The mastery of a section is the average mastery of the exercises related to the respective concepts of the section. At the end of an exercise, the user can provide feedback, i.e. her confidence with a solution, whether she liked the exercise, if she is happy with the solution, and satisfied with the result (cf. [MAF+01]).

A user's mastery knowledge is represented by three values drawn from *Bloom's taxonomy*[Blo56], namely knowledge, comprehension, and application. Depending on the type of user interaction, different updates of the values are realized. For example, after reading a concept, its *knowledge-value* is updated, after reading an example for a concept, its *comprehension-value* is updated, and after solving an exercise, the *application-value* of the concept is updated. Additionally, a justification for each value is stored in form of a pointer to a history component, called HistoryAtoms [MBA+01], responsible for the changes. HistoryAtoms contain information such as the ID of the page the user read, the reading time, and the success rate of exercises on the page. The current granularity of HistoryAtoms is the page level. However, attempts have been taken to trace the user's attention and reading time on a more detailed level (cf. [MBA+01] and [MAF+01]).

To sum up, LECTORA LIGHT and ACTIVEMATH are both based on a semantic representation language, which specifies content and its structure. Both approaches include user modeling, based on either direct user feedback or assumptions of the system. However, ACTIVEMATH only collects user information based on explicit user feedback or questionnaires as well as tracing the user's learning progress. In contrast, LECTORA automatically detects user information from the user's collection of documents. Furthermore, LECTORA LIGHT infers user practices based on a formal representation and thus may offer better explanation and justification of the user modeling process than ACTIVEMATH. However, the initialization of user models remains a problem for both systems. An additional distinction lies in the target users of both systems: While ACTIVEMATH is a learning environment for students, LECTORA LIGHT also aims at researchers and scholars.

The improved prototype LECTORA intends to identify communities of users. First attempts of adapting the types of examples and exercises based on a user's field are taken in ACTIVEMATH. However, LECTORA provides further improved adaptation based on information on a user's community, e.g. his current lecture. For example, LECTORA bases the selection of examples on the assessments by other members of the user's communities. In addition, having identified different mathematical communities, LECTORA may offer further services such as visualizing, describing, and comparing communities. Moreover, by allowing the student Normen to explicitly subscribe to a COP, e.g. the student's lecture, and by knowing the COP's descriptive practices, LECTORA can automatically initialize Normen's profile by the set of practices that describe the respective lecture. Based on the initial profile, LECTORA supports Normen to understand the specific focuses of a lecture, learn about alternative notations that are used, and to quickly identify new assumptions a proof is based on. In addition, the system considers Normen's feedback to modify his profile.

Notation Selection Elena Smirnova and Stephen Watt describe a software tool [Not07] that provides output according to specified notation preferences. They implemented a notation selector by a graphical user interface, in which users manually specify their notation preferences. Based on the user's selections, an XSLT stylesheet is generated, which is used to convert XML-based conceptually-oriented mathematical documents, e.g. in Content MATHML format, into a notation-oriented format, e.g. in Presentation MATHML (cf. [SW06]). The kill of this approach lies in the manual selection of notations, which constitutes an enormous additional effort by the user and prohibits a dynamic adaptation based on the user's interactions. Furthermore, the context of notation, which might influence the user's prefer notations, is not considered. Although an initial selection of preferred notations may be considered, LECTORA LIGHT intends to automatically detect and dynamically modify notation preferences from the user's document collection. In addition, LECTORA aims at providing further services, such as using notation preferences to identify a user's community and based on that to improve search rankings and recommendations.

2.3 Annotation and Browsing Tools

Connotea [con07] is a reference management service that focuses on the improvement of scientific information access. The system applies social computing methods by offering annotations of other users to assess existent and new scientific information more efficiently. Connotea is based on the semantic annotations frameworks Annotea proposed by the W3C [W3C07]. Annotea implements simple annotations, i.e. comments, notes, and explanations, by using XPointer [XPo07] for locating annotations within XML-based documents. LECTORA may build upon the framework used by Connotea, however, annotations in LECTORA have to be far more sophisticated. In particular, they are required to be semantically structured to be machine-understandable and subject to automatic inferences.

mSpace [msSO+05] is a personalized information presenter that provides facetted browsing to improve the ordering of searches for information based on the user's associations and his browsing behavior.

Didactic Storyboarding: Storyboarding is an approach for modeling the user's interaction in a learning system: An information space, predefined by the authors, is given, in which users can freely navigate according to their needs. Storyboarding is implemented by annotated graphs in which nodes represent scenes or episodes, edges represent the transitions, and annotations allow to indicate the author's intention. The Storyboarding implementation by Klaus Jantke and Rainer Knauf [JK05] uses annotated graphs in which nodes link to specific educational content, e.g. modules of a lecture, examples, assignments, and quizzes, and in which edges specify the transitions between them. In addition, annotations, e.g. assessments and associations of students, can be added to nodes and edges. The graph is implemented in MS Office Visio [Vis07]. This approach has two benefits: First, assessments and associations are visualized, and, secondly, a collaborative and dynamic development of graphs can be implemented. However, the storyboarding approach presented needs to be extended in order to be useful LECTORA LIGHT. For example, the current implementation does not provide the required granularity of content. Furthermore, annotations should be formalized to allow inferences on the provided feedback.

To summarize, Figure 3 visualizes the related work for the implementation of LECTORA LIGHT. For example, research in the field of living documents, such as ACTIVEMATH, relates to the user modeling and presentation module of LECTORA LIGHT. In contrast, the information retrieval module benefits from existing work on recommender systems as well as browsing tools, such as mSpace. The reference manager Connotea relates to the annotation feature of LECTORA LIGHT. However, the previously discussed related work is most likely to support the implementation of the light version of LECTORA, while the improved version of LECTORA augments previous approaches by modeling and using CoPs for improving its services. Thus, the research on social computing and collaborative recommendations is used in the specification phase of the community module of LECTORA, but leaves much room for further discussions and evaluations.

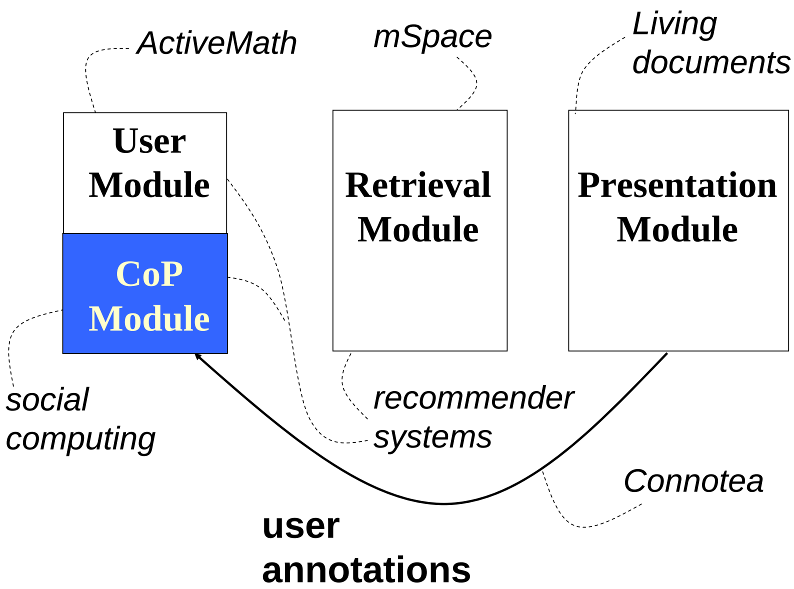

Figure 3: Related Work of LECTORA LIGHT.

3 Mathematical Practice

As mentioned earlier, this work applies the concept of Communities of Practice (CoP) to the research on personalized information access in the domain of mathematics. Therefore, this section introduces the theory of CoP, discusses several mathematical practices, and gives concrete examples of notation practices that help to distinguish mathematical communities.

3.1 Communities of Practice

Just like other research domains, mathematics can be divided into several heterogeneous communities. The members of these communities share certain background knowledge and common tasks. While interacting, they collectively learn to improve achieving these tasks, which results in commonly agreed on practices, i.e. procedures, notations, and workflows. This section introduces Lave and Wenger's theory of Communities of Practice (CoPs) [LW91] and emphasizes its applicability to describe mathematical subcultures.

In his work, Wenger [Wen05] emphasizes four main aspects of practices that determine the internal dynamics of a CoP, particularly practices as meaning, community, learning, and boundary:

Practice as meaning Wenger [Wen05] discusses the negotiation of meaning as well as two interoperating processes, namely participation (action and connection) and reification (objectification and evaluation). In this sense, *mathematics can be described as a continuous process of creation and negotiation of meaning*, in which reification transfers experiences into mathematical objects: According to

Philip Kitcher [Kit88], the development of mathematics can be seen as a stepwise process from generalizations and observations to their symbolic substitutes. Kitcher underlines the dynamics in mathematics, i.e. the creating, revising, and dismissing of mathematical knowledge, as well as the process of abstracting experiences to gain symbolic substitutes, i.e. by reification (cf. [Hei00]).

Practice as community Wenger [Wen05] stresses that in order to form a CoP, participants have to be mutually engaged in actions whose meaning they negotiate with one another. This also applies to mathematical communities, in which, based on different background knowledge and tasks, participants interact regularly in order to achieve a shared task. Doing so, members of a community may be involved in various enterprises, i.e. the communal creation of new mathematical knowledge (proving of a new theory) or the coordination and sense-making of existing mathematical knowledge (including the disproving of an accepted theory). Furthermore, mathematicians strongly cooperate, need to form common understandings and depend on each other's results. However, doing mathematics remains an intuitive, creative, and subjective process. Thus, *participating in a mathematical community is a personal and communal activity*: Individual knowledge, experience and preferences have to be considered as much as the communally agreed on communication rules and enterprises.

Practice as learning Wenger [Wen05] argues that *practices change over time* and that a CoP is therefore an emergent structure. Accordingly, the membership of mathematicians to one or more communities varies as they forget, remember, or change their perspective and participation as well as gather new experience that may influence their understandings and beliefs. When changing their practices, mathematicians terminate their participation in one subculture and start involving in another one simultaneously. Whenever new members join a subculture, discontinuities are provoked, i.e. older members are forced to reveal or revise their perspectives and knowledge. Thus, *learning takes place on an individual as well as communal level*: Individuals learn by engaging and contributing to the practice of a community; while communities learn by refining their practices and ensuring the participation of new members.

Practice as boundary This relates to the boundaries that CoPs create, including the types of connections that may create bridges across CoPs. Although, as argued before, subcultures are emergent structures with rather *fuzzy boundaries*, certain characteristics or practices can be identified to distinguish them. Andrea and Michael Kohlhase [KK06] argue that reification as well as participation are inscribed into mathematical documents. Thus, *practices (e.g. notations) of a* CoP *can be detected in a collection of documents*, which constitute boundary objects [Wen05].

In conclusion, mathematics constitutes of several communities of mathematical practices. By participating in these communities, meanings of mathematical objects and practices are continuously discussed and eventually changed. In consequence, CoPs are emergent structures with fuzzy boundaries: Although they can be distinguished by practices, these practices may change over time and, thus, also the members and characteristics of a community may change. Knowing about practices and their dynamics is therefore an important aspect in order to support individual learning and understanding as well as to provide community-based search features and adaptations. The next section discusses mathematical practices and deduce requirements for LECTORA.

3.2 Mathematical Practices in the Literature

The following section includes a discussion of mathematical practices according to Bettina Heintz [Hei00] and George Polya [Pol73].

Proving is one of the widely known practices of mathematicians. According to Heintz [Hei00], proofs are highly standardized communication procedures that support an intersubjective communication between mathematicians. However, proofs need to be understood in order to conceive the mathematical idea behind. Practices such as *axiomatizing, formalizing, choosing specific notations, explicating the terms and definitions* used may help to illustrate the development of the proof. However, there are two problems: First of all, practices may change over time. Secondly, *proofs are mostly not complete and include gaps*, i.e. not all information can be explicitly given in the proof. It is assumed that this missing information is added by the reader. Thus, a *common background is required* in order to make sense of communication practices and to fill gaps of the proof. According to Heintz [Hei00], a common background is *created by direct non-formalized communication*.

Communal Acceptance: Heintz [Hei00] highlights that the **acceptance of a proof** by the mathematical community, is an important mathematical practice. This is underlined by the citation by Yuri I. Manin[2] [Man77] according to whom "a proof only becomes a proof after the social act of accepting it as a proof". Heintz [Hei00] points out that mathematicians experience an extreme pressure evaluating whether e.g. they can use a mathematical result in a journal for their own research or not. In order to find out, they need to *know about the reputation of the author, the prestige of the journal or the usage of the results by other mathematicians*. Since mathematicians highly depend on the work of their colleagues, the *correctness of results* is of high importance. Many times results are used without checking them personally. In consequence, errors propagate and are sometimes hard to detect. In order to ensure correct results, two main practices have been developed to ensure the correctness of publications: The referee process and the sending of preprints to colleagues. However, Heintz [Hei00] stresses that this is not a guarantee for the correctness of mathematical results. They still need to be *understood, accepted and checked* by the mathematical community as a whole.

Cooperating: Another practice in mathematics mentioned by Heintz [Hei00] is **cooperating**. According to Heintz, *mathematical cooperations* are voluntarily, arise spontaneously to solve a certain problem, terminate after a solution for this problem has been found, and comprise a small number of participants, who communicate informally. Accordingly, CoPs are informal and emergent structures that only exists for a limited period of time. This argumentation points out the applicability of the CoP concept to the domain of mathematics.

Problem Solving: Based on his guideline for solving mathematical problems, Polya [Pol73] distinguishes several mathematical practices during the problem solving process that differ among communities and their members. These are *finding analogous problems, introducing an auxiliary element, solving an auxiliary problem, drawing a figure, using generalization or induction, reusing previous solved problems as well as using meaningful notations, heuristics, analysis, or specialization*.

Notations: In his work, Polya [Pol73] presents several examples for mathematical practices such as best practices for choosing meaningful notations, which he requires to be unambiguous, pregnant, and easy to remember: For example, it is essential to use meaningful initials as symbols, such as a V for Volume or an r for radius. Furthermore, a notation is helpful when the order and the connection of the sign suggest the order and the connections of the objects, e.g. by using letters at the beginning of the alphabet (a, b, c) for given quantities or constants and letters at the end of the alphabet (x, y, z) for unknown quantities or variables. Alternatively, a best practice is to denote objects that belong to the same category with letters of the same alphabet, i.e. Roman capitals (A, B, C) for points, small Roman letters as (a, b, c) for lines, Greek letters (α, β, γ) for angles. Although the best practices above are commonly agreed on guidelines for various authors, there might be variation among mathematical communities. Thus, the comprehensibility of a publication can be increased by knowing which presentation and naming conventions are preferred and most easily understood by its potential readers.

[2] Yuri Ivanovich Manin is a Russian mathematician, known for his work in algebraic geometry and diophantine geometry, and many expository works ranging from mathematical logic to theoretical physics.

To summarize, the acceptance of content and navigation in new research fields can be improved by providing information on the reputation of the author, the prestige of the journal or the usage of the results by other mathematicians. Furthermore, illustrating the development of a mathematical idea without gaps and adding supplementary information, such as explanation of notation and terms, definitions, explications of axiomatization and formalization procedures, the comprehensibility of proofs can be improved. In addition, the ability to solve problems can be supported by e.g. referring to known content, choosing comprehensible notations, generalizing and specializing content, providing additional content such as more examples, figures, explanations as well as explicating assumptions.

3.3 Mathematical Notation Practices

After introducing several mathematical practices in the previous section, this section emphasizes the presentation of mathematical notations among different communities. According to Stephen Watt and Elena Smirnova [SW06], the existence of numerous notations is caused by a number of reasons, namely the mathematical context, the area of application, national and cultural conventions, and the level of sophistication. The following examples illustrates the importance of adapting documents to the notation preferences of users.

An Example for the level of sophistication: While $a \div b$ is mostly used in elementary school, $\frac{a}{b}$ is used in higher education. (cf. [SW06])

An Example for the Area of Application: To introduce an *imaginary unit of a complex number*, a mathematician uses the symbol *i* to specify a complex number as $a + bi$. In contrast, an electrical engineer, uses *j* instead of *i* to avoid confusion with the symbol *I* for electric current.

An Example for National Conventions: A German researcher specifies the *binomial coefficient* by the notation $\binom{n}{k}$. In contrast, a Russian researcher uses the notation C_n^k, while a French researcher will most likely use C_k^n.

In conclusion, we can identify several notation preferences as well as factors that may indicated the preference of a reader. For example, the origin, e.g. Russia, or national convention allows a system to infer a presentation of a binomial coefficient, e.g. C_n^k. It is assumed that the automatic adaptation of notation presentation increases the comprehensibility and, even more important, also ensures the *consistency* and *reusability* of specifications between research communities. For example, when entering content in a knowledge base, a researcher is more likely to choose a notation he is familiar with and/or that relates to his area of application. Hence, in order to reuse content, the notation itself needs to be represented separately from its presentation. Furthermore, the mathematical context of the notation needs to be specialized. Based on that, a consistent notation can be chosen according to the current context of the document and be linked to an alternative presentation and the respective mathematical context. However, a Russian researchers that is strongly collaborating with a German Research Lab might as well adapt to their presentation and eventually also prefer it. In consequence, the reader needs to e.g. consider the user's satisfaction with the automatic translation, and use his feedback to continuously update the user model.

4 LECTORA — System Description

LECTORA (Spanish for female reader) is an interactive and collaborative reader for researchers and students that supports a more efficient access to mathematical documents. Based on the semantic representation format OMDOC [Koh06], the system adapts content, structure, and appearance of documents. The first phase of this work, focuses on the inferring of user practices based on different user information and the implementation of LECTORA LIGHT, a practice-oriented reader. The second phase focuses on the modeling of Communities of Practice, the attribution of users to communities as well as the implementation of LECTORA, a community-aware reader.

4.1 Practice-Oriented Phase

The first phase is a preparation phase in which practices of researchers, scholars, and students are formalized and user-specific practices are modeled. It aims at building a light version of the reader that infers a user's practices from different user information, i.e. browsing behavior, document collection, and feedback, based on a formal representation model of practices. Since these practices describe a specific user, they are stored in the user model and are used to provide user-centered and practice-oriented services. The first phase is completed by an educational case study and the implementation of LECTORA LIGHT.

Formalization of Practices Before giving details on the formalization of practices, three different approaches for collecting information on users shall be introduced: (1) Some user practices may be collected by tracing the user, i.e. by monitoring his browsing behaviors and interactions. (2) Alternatively, explicit user feedback, i.e. by means of questionnaires or ratings, can be analyzed. (3) This works assumes that especially mathematical practices, i.e. choosing a specific notations, are inscribed into scientific documents. Hence, user practices may also be detected by analyzing the user's document collection, i.e. the set of documents he writes and reads. The more structured the respective format of the documents, the more practices can be identified. Thus, LECTORA LIGHT is based on a semantic representation format for mathematical documents, i.e. Open Mathematical Document OMDOC [Koh06], which e.g. facilitates the automated identification of different notation presentations, content, or order of content (cf. section 2). Figure 4 presents an example in which two practices are automatically detected from a document, namely the presentation of a binomial coefficient and the referencing of related work. A third practice is gained by analyzing the annotation of the user Immanuel, namely his interest or the relevance of the document.

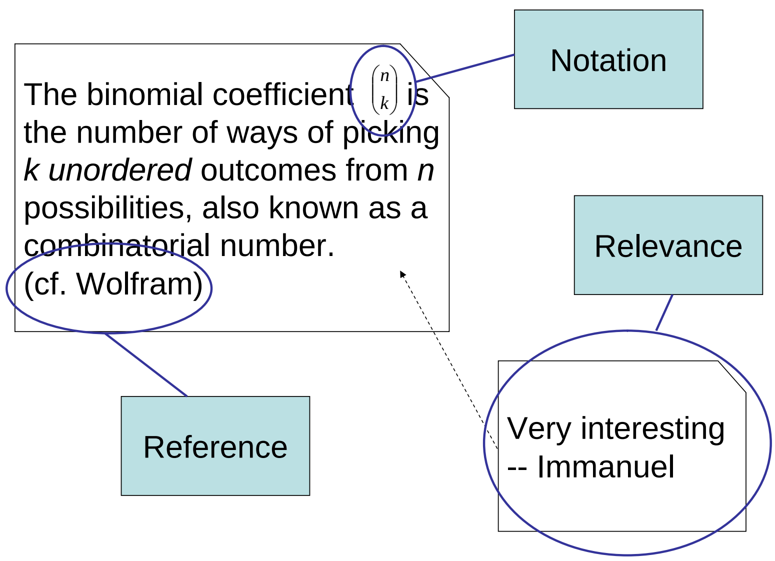

Figure 4: Extracting Practices from Documents.

In order to enable LECTORA LIGHT to infer different practices based on e.g. the user's feedback or documents, practices need to formalized and inference rules need to be represented in the system. Figure 5 provides an example of how practices can be formalized, i.e. by identifying *practice concepts*, such as 'notation', and *sub-concepts*, such as the 'notation of a binomial coefficient' as well as *practice instances*, such as alternative representations of a binomial coefficient. In addition, *relations* among practices need to be identified. For example, the origin of a user might influence the notations he prefers. A concept and an instance of the concept, e.g. the notation of a binomial coefficient and the presentation

$\binom{n}{k}$, constitute a *practice object*. Practice objects and their interrelations are represented in a CoP *Meta Model*. In LECTORA LIGHT, this CoP Meta Model is used to map the user's relevant characteristics and behaviors to practice objects, which are stored in the respective user model. Hence, a user model contains practice objects and their interrelations.

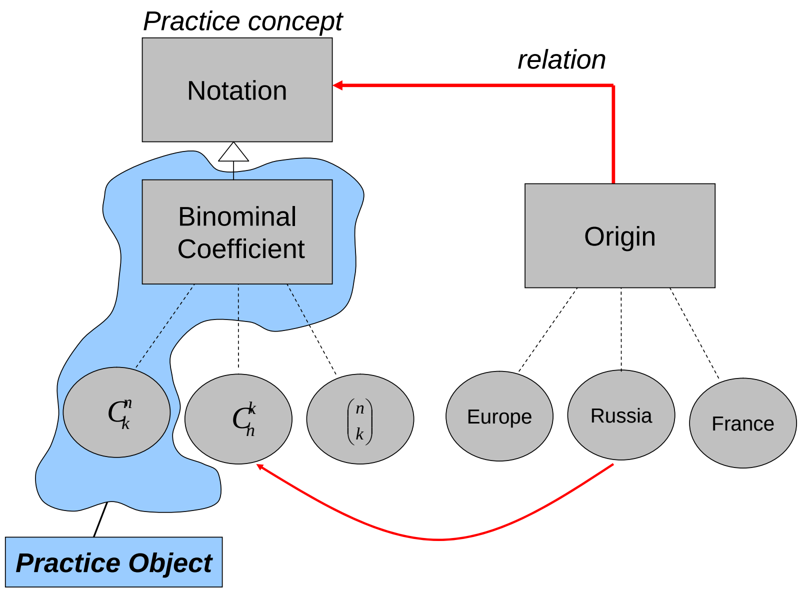

Figure 5: An Example of Formalizing Practices.

Practice-Oriented Services Figure 6 presents the three modules of LECTORA LIGHT: (1) a practice module including the user models, CoP meta model, and inference rules, (2) a retrieval model providing recommendation and search rankings based on the user model, and (3) a presentation module which adapts the content, structure, and appearance of documents.

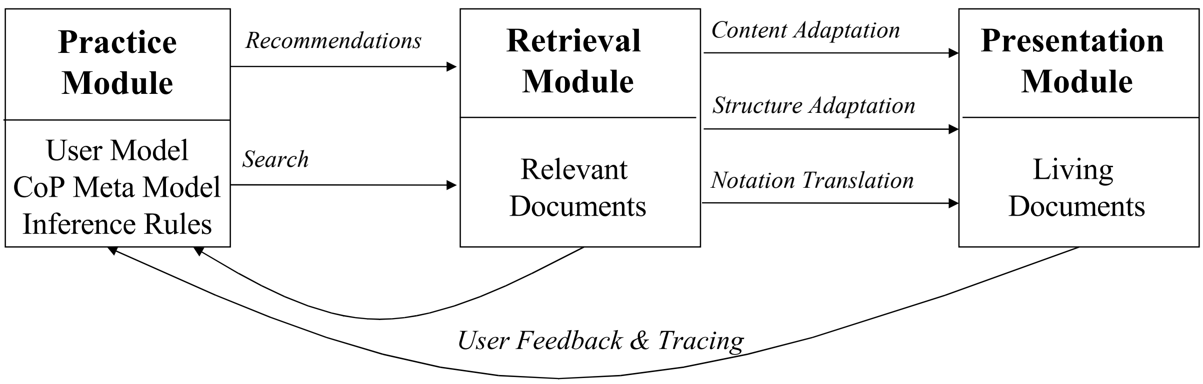

Figure 6: Practice-Oriented Services of LECTORA LIGHT.

After the user logs into the system, LECTORA LIGHT retrieves the latest snapshot of his user model, i.e. a set of practice objects that the system has inferred based on the user's explicit feedback, behaviors, and documents. The user's feedback includes his assessment of e.g. the quality and usability of recommendations made by LECTORA LIGHT, the ranking of search results, and adaptation of documents. It also contains his judgment e.g. upon the relevance, significance or correctness of documents. This information is used to infer e.g. the user's preferred presentation of notation, his preferred structuring of mathematical content, as well as his preferred research topic and interest. For example, a user could indicate that he dislikes the current notation and ask the system to choose an alternative one instead. The user's behavior include his browsing behavior and his interactions with other users via the system. Further information of the user, in particular his organization affiliation, background, education, work experience, and his association to other researcher groups could also be considered to infer further practices. For example, a user's membership to a specific organization may allow LECTORA LIGHT to draw conclusion on the user's background. Or alternatively, knowing about a user culture and origin, LECTORA could assume specific notation practices.

The more information LECTORA LIGHT gathers about the user, the better the usability of recommendations and the quality of adaptations becomes. In consequence, a main issue is to solve the problem of initiating the user model of new users, henceforth referred to as *newcomers problem*. Part of the initial user information is gathered by asking new users to fill out a questionnaire in which they indicate the organization they work with, their mathematical background, the education they have passed, as well as their collaborations. However, the most valuable information is captured by monitoring the user's interaction with the system, e.g. by storing information on which documents the user has read as well as capturing his annotations and assessments. To receive this valuable information, users have to be assured that LECTORA LIGHT continuously learns to improve and adapt it's service, and that for this reason their investments in the system will eventually pay of.

After identifying a user's practices, LECTORA LIGHT offers two option in the retrieval module: automatic recommendations and manual search. For example, LECTORA LIGHT recommends documents based on the user's (current) interests as well as further aspects such as the user's context and information on the content itself. While recommender system have already addressed the problem of personalized information retrieval, the strength of LECTORA LIGHT lies in her ability to explicate the retrieval rules. For example, she can point out the practices of the individual user and the respective rules that they triggered in order to explain to the user why a certain document was recommended.

After finding a relevant document, LECTORA LIGHT provides user's with individual views on the document. By considering the user's preferred structure, notations, and his interest LECTORA LIGHT automatically adapts the document. For example, LECTORA LIGHT automatically selects notations according to the user's preferred presentation. Doing so, LECTORA LIGHT also ensures the consistent use of notations in documents. However, automatic adaptation is not always the best method to learn about a topic. Strictly speaking, it may even have negative impact on the users' learning experience since they are no longer required to cross-link various notations. Hence, instead of automatically translating notations, LECTORA LIGHT optionally offers a listing of alternative notations and their respective context. In addition, LECTORA LIGHT also adapts the structure and content of documents according to the user's preferences. For example, users might be more likely to understand a document if less complex examples are added or already known information is linked to new one. Alternatively, user may have certain layout preferences, such as the highlighting of definitions. However, the adaptation also depends on the purpose of using LECTORA LIGHT: For example, students might be forced to read a whole document and to identify known problems that help solving new examples. In contrast, a researcher focuses on getting the rough idea of a paper without understanding details of a proof and may be satisfied with an extract of the document.

4.2 Community-Aware Phase

The previous section presented several practice-oriented services of LECTORA LIGHT. However, the innovation of this work lies in the knowledge about a user's communities of practice. Thus, the second phase focuses on the modeling of COPs, the application of COPs to provide community-aware services as well as a case study and the implementation of an improved version of the prototype, i.e. LECTORA.

Modeling of Communities Of Practice In section 1.2, CoPs have been defined as practice-oriented groups of people who share similar tasks and who, share specific practice in order to solve the task. The sets of shared practices of its members describe a respective CoP. In LECTORA, CoPs are represented by CoP models. Figure 7 visualizes the analogy between user modeling and CoP modeling. In LECTORA LIGHT the CoP meta model, i.e. the formal representation of practice objects and their interrelations, was used to infer a user model from the user's behaviors, annotations, and documents. In Figure 7 the user Achim and the user Michael are represented by a user model that contains their individual practices, which were inferred from their documents, annotations, and behavior. Hence, user models are sets of user-specific practices and represent users in the system. Accordingly, CoPs are sets of practice objects that describe and represent CoPs in the real world. In Figure 7, the KWARC[3] [KWA07] and DFKI[4][DFK07] CoP are represented by a respective CoP model that contains sets of practices that describe both CoPs.

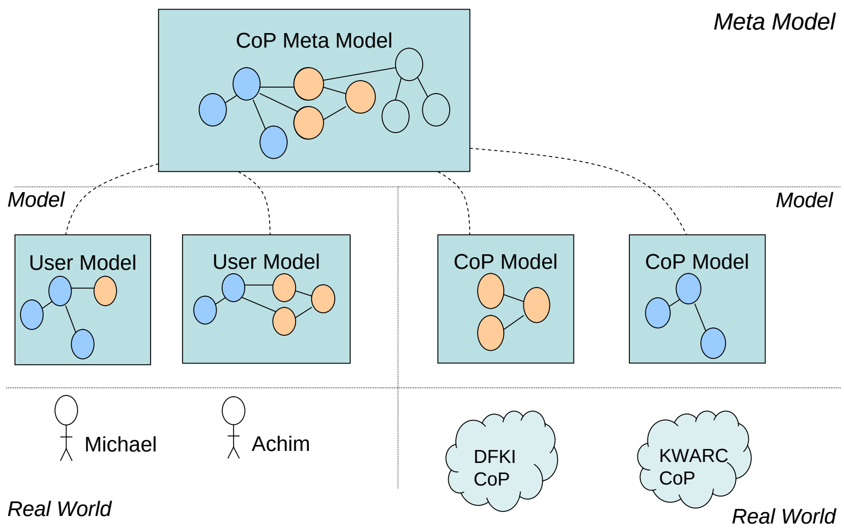

Figure 7: User Model, CoP Model, and CoP Meta Model.

There are two methods for modeling CoPs: an *extensional approach* and an *intensional approach*, as visualized in Figure 8. The extensional approach models CoPs based on the shared practices of its members: The identification of clusters of members allows to identify their shared practices and to describe their CoP. In contrast, the intensional approach requires the specification of characteristic practices of a specific CoP and based on that allows for the identification of members. Accordingly, phase I of this work provides several user models, i.e. sets of practice objects that describe specific users, for an intensional approach: By means of cluster analysis an optimal number of clusters can be identified based on the similarity of user models. For example, in Figure 9 four clusters have been identified that represent four CoPs, namely the CoP of the KWARC group, DFKI members, Wikipedians, and Statisticians. The KWARC group shares similar tasks in the field of knowledge representation. Some of its members also share tasks with other CoPs. For example, the KWARCie Christoph works on the implementation of a semantic Wiki and thus relates to the Wikipedian CoP. In contrast, another KWARCie Normen works on management of change and thus relates to Achim, a member of the DFKI, who focus on managing variants. After identifying clusters, sets of practices need to be identified that describe them. For ex-

[3] KWARC is the 'Knowledge Adaptation and Reasoning for Content' research group at the Jacobs University Bremen and the author's working group, i.e. one of her communities of practice.

[4] DFKI is the 'German research insitute for artificial intelligence' and one of the close collaborators of the KWARC group. This makes the DFKI an additional CoP for most KWARC members.

ample, the members of the KWARC CoP share the practice of using a semantic representation format and communicating via a Wiki. Practices may also be shared between CoPs. For example, Achim bases his work on the same representation format and, consequently, shares this practice with the members of the KWARC CoP. However, the automatically identified sets of practices need to be evaluate, e.g. manually by users, in order to detect mavericks. For example, based on the cluster analysis sets of similar practice among the KWARCies have been identified. However, some practices are not relevant for describing the CoP, e.g. going shopping or sleeping, and need to be distinguished from the applicable ones.

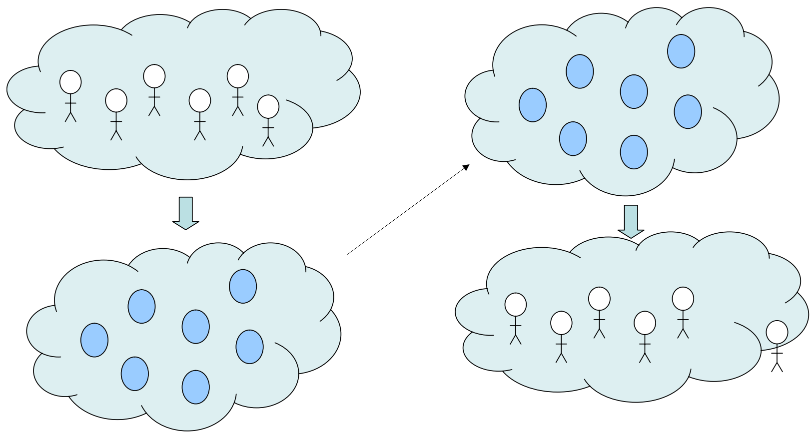

Figure 8: Intensional and Extensional Approach for Modeling CoPs.

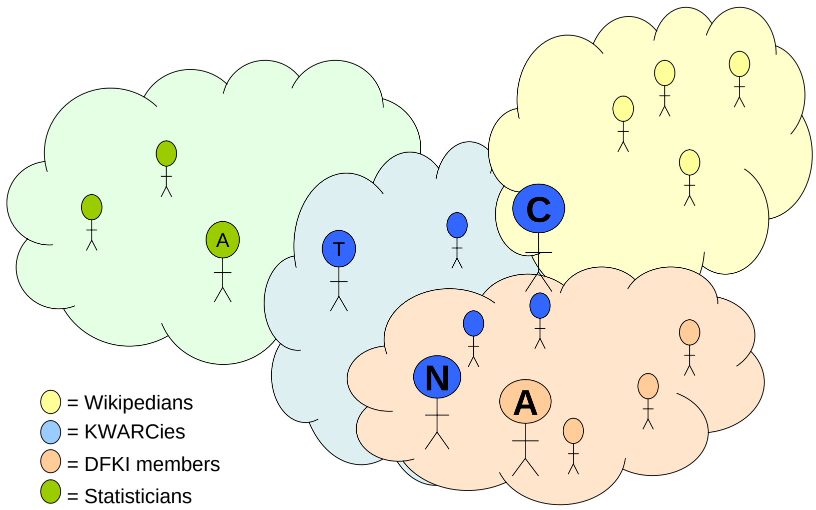

Figure 9: Clusters of User Model.

Having identified sets of practices by an extensional approach, the description of CoPs by these characteristic sets may be use to attribute users to a specific CoP. For example, depending on the similarity between the CoP's model and the user model, a user is identified to be a full or a partly member of this CoP. Thus, LECTORA applies both an extensional approach to find sets of practice objects that describe a CoP and an intensional approach to use the description of a CoP to find its members.

Community-Aware Services Based on the explicit description of CoPs, LECTORA provides means to solve the newcomer problem LECTORA LIGHT is still facing (cf. section 4.1). For example, after a new user explicitly chooses one or more CoPs, LECTORA can initialize his user model by retrieving the relevant practices from the respective description of his CoPs. The descriptions of CoPs are also useful in a learning scenario. For example, in Figure 10 Achim has been identified as a partly member of the KWARC CoP since he shares two of the KWARC CoP's practices, i.e. practice 1 and 2. However, in order to become a full member of the KWARC CoP he still needs to learn practice 3. LECTORA is able to identify that Achim still needs to learn this specific practice, e.g. the representation format used by the group. The system also supports Achim to acquire the specific practice, e.g. by recommending a tutorial that explains the characteristics and use of the format.

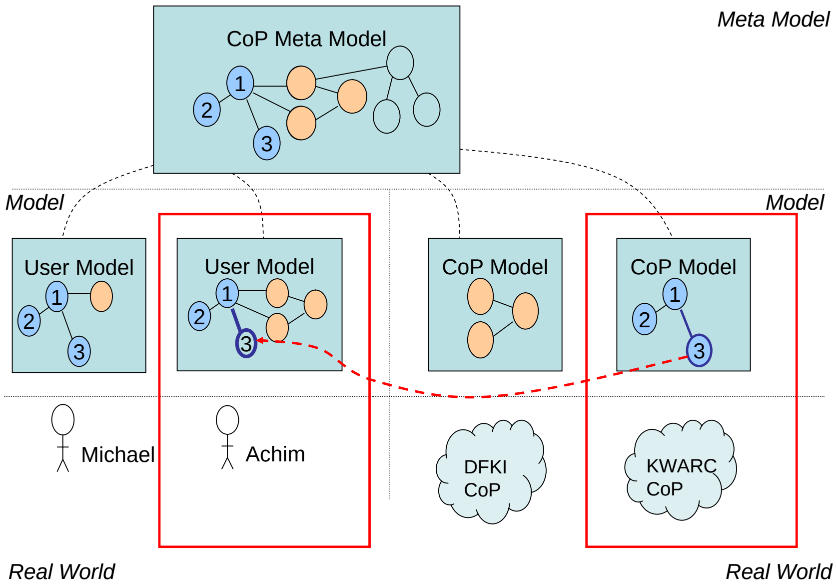

Figure 10: Learning Scenario

An alternative service LECTORA offers is a community-based ranking of search results. For example, in Figure 11 the two CoPs KWARC and DFKI have rated several documents differently: The KWARC CoP has ranked document K, W, A, R, and C to be highly relevant, while the DFKI CoP specified document D, F, K, and I to be interesting. Let's assume that Achim is a full member of the DFKI and partly member of the KWARC, while Michael is a full member of KWARC but only partly member of the DFKI. By considering the users membership as well as its strength, LECTORA generates personalized recommendation for both researchers: LECTORA recommends Achim to read documents D, K, and W, while the system recommends Michael to have a look at documents K, W, and F.

A further innovation of LECTORA, as opposed to existing approaches, such as implemented in several recommendation systems, is the consideration of contextual information, i.e. time, place, and groups of users an information is shared with. The context may influence the relevant user practices LECTORA should consider and may have an impact on the current communities of practice LECTORA identifies. For example, practices change over time. Hence, while studying mathematics, Normen may be used to the the symbol i for an imaginary unit. However, his notation preferences may change if he decides to start his Ph.D. in the area of electrical engineering since he will eventually end up using j instead (cf. section

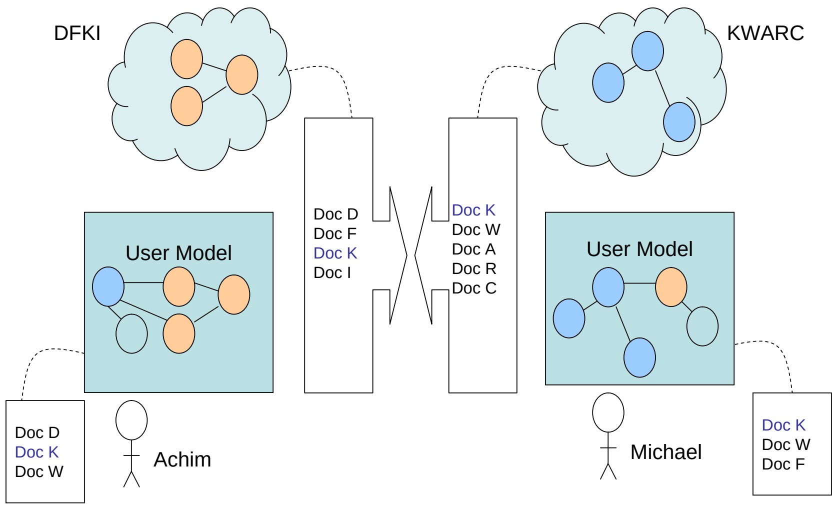

Figure 11: Providing Personalized Recommendations

3.3). In contrast, the relevance of documents may also depend on the current context of the user. For example, the researcher Immanuel currently involved in the MKM-IG (Interest group of Mathematical Knowledge Management) [MKM07] will assess the relevance of documents according to the interest of this group. However, when assessing the relevant literature for his students, he will come up with different documents.

5 Case Studies and Prototypes

I will undertake two educational case studies to evaluate the applicability of the proposed reader, in particular LECTORA LIGHT and LECTORA.

5.1 Educational Field Study: LECTORA LIGHT.

The first field study involves students, teaching assistants, and the professor of the General Computer Science lecture (GenCS). One goal is to analyze the students' problem solving behavior, practices they perform to recapitulate a lecture, to prepare for an exam, or to solve an assignment. Thus, several information on students has to be detected: For example, their assessments and association of the professor's lecture material, quizzes, and assignments in terms of their degree of difficulty, comprehensibility, and/ or interest is analyzed. Further information on the students' approaches of recapitulating the lecture, solving assignments, and preparing for quizzes and finals, has to be captured. A survey is one possible alternative to gather relevant information, e.g. by means of a questionnaire that asks volunteers to associate examples that helped them solving a assignment or asking them to rank the difficulties of the assignments. Alternatively, students could be encouraged to rank the assignments when uploading their solutions via the GRADER System, an interface by which students can submit their programming assignments of the GenCS lecture. Another approach is the Storyboarding implementation by Klaus Jantke and Rainer Knauf [JK05] that allows learners to annotate educational content and the relations between it (cf. section 2.2.

In addition to the practices of students, the practices of the professor shall be considered: Assuming that his practices are inscribed in his lecture material, slides, notes, and assignments can be used to first identify practice and second to come up with approaches to automatically detect practices in other lecture material.

Based on the case study, practices of students and the professor are formalized in a CoP meta model. Furthermore, LECTORA LIGHT provides means to retrieve and present the materials and assignment of the GenCS lecture. Furthermore, the system models practices of its users by analyzing their browsing behavior and feedback as well as provides means to automatically detect practices in a document. Based on this information, recommendation, search rankings, and adaptation of educational material is provided.

Finally, the result of the case study and services of the prototype are compared to existing approaches, such as ACTIVEMATH (cf. section 2.2). Furthermore, the assumptions whether practice are inscribed into documents and whether a semantic representation facilitates the automatically detection of practices, are verified.

5.2 Educational Case Study: LECTORA

The second case study augments the GenCS scenario. The improved version of the prototype models CoPs based on the students user models. The professor's practices, automatically detected from his documents, are visualized and can help students to understand the expectation, assumptions, and approaches that differ from the ones they know. Based on the description of the lecture, LECTORA detects if a student lacks certain practices and, if required, recommends him to look at additional material that helps him to acquire a certain practice. Furthermore, the visualization of how students approach the lecture offers valuable advice for other students, e.g. by quickly finding examples that are helpful or associating the assignments with a topic in the lecture. LECTORA also identifies sub-communities within the lecture and recommend students to collaborate with students with similar practices, e.g. the preferred time of solving an assignment. In addition, the students' assessment upon the lecture provides valuable feedback for the professor, which is not restricted to a general questionnaire at the end of the lecture but rather directly refers to specific problems in the GenCS lecture.

The services of the improved prototype are compared to LECTORA LIGHT. Furthermore, the assumptions whether CoPs can be identified by clusters of user practices, and whether the descriptive theory of CoP can be used to predict the preferences of users, are verified.

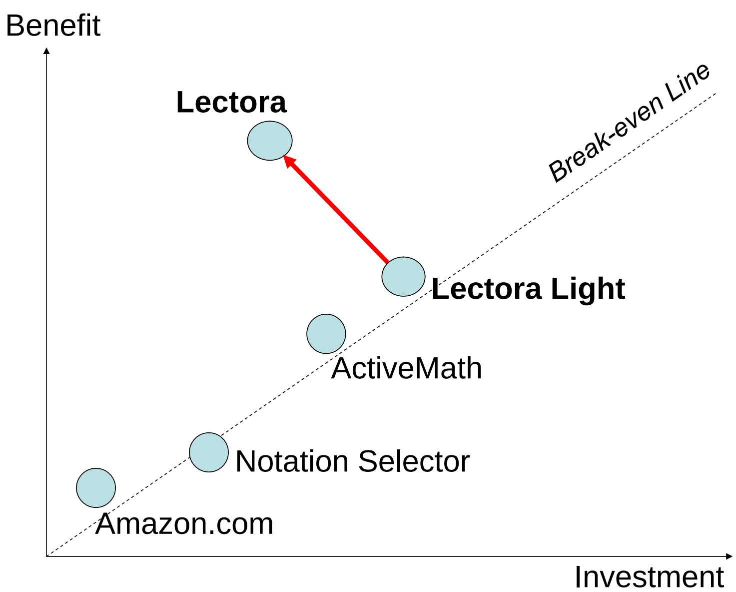

Figure 12: The Vision.

Figure 12 shows the vision of this work. It shall be emphasized that the placing of related work is not based on a representative survey but rather shall underline the expected innovation of this work: In Figure 12, Amazon.com [LSY03] is placed in the lower left hand corner since the system requires minimal additional efforts of users besides the investment of purchasing items. However, based on the shopping behavior of user and current shopping cart, the approach does not consider further associations between user and may only provide simple recommendation, that e.g. neglect new or unusual items. The notations selector [Not07] by [SW06] facilitates the automatic translation of notation based on the user's preference. Hence, it provides a fairly high benefit to the user. However, users have to manually enter their notation preferences, which requires an enormous additional investment and prohibits a dynamic adaptation based on the user's interactions. The user-centered learning environment ACTIVEMATH generates personalized educational courses based on user models. Thus, the benefit for the user is relatively high. However, the system requires the participation of a user and his feedback in form of questionnaires and ranking. In contrast, LECTORA LIGHT aims at automatically detecting user information from a collection of documents and, by doing so, reduces the investment of users. On the other hand, LECTORA LIGHT requires more detailed feedback of the user, i.e. his assessment of the relevance and quality, and requires a formal representation of the user's annotation. Unless a user-friendly editor is provided, users have to specify the semantic structure of their annotation instead of being allowed to enter informal text. In consequence, LECTORA LIGHT requires more additional efforts from the user than ACTIVEMATH. On the other hand, LECTORA LIGHT augments ACTIVEMATH by offering additional services, such as explanations and justifications of the user modeling and adaptation process, and thus, provides a greater benefit. Above all, LECTORA LIGHT is a pre-study prototype for the improved system LECTORA, which based on the findings in the first phase provides innovated solutions to e.g. the newcomer problem and thus, tremendously reduces the initial investment of users. Moreover, LECTORA identifies a user's CoPs and uses these community-related information to offer augmented service such as the ranking of search results and personalized recommendations. In addition, LECTORA describes, visualizes, and compares CoPs and facilitates users to adapt to new communities as well as learn about new practices. To sum up, LECTORA LIGHT is expected to provide greater benefits but also an increased investment of users. In contrast, LECTORA is expected to tremendously reduce the initial investment and further efforts thanks to the modeling of CoPs, which also allows the system to offer a much greater benefit.

6 Workplan

Requirements Gathering and Analysis. I spent the first six month reading up on the concept of Communities of Practice by [Wen05], general and mathematical knowledge management, information retrieval, recommender systems, statistical methods such as cluster analysis, and further related work, i.e. ACTIVEMATH or mSpace. Thus, I broadened my understandings of the current state of the art in personalized information access and was able to identify requirements for a the modeling of practices and CoPs, as well as potential high-value service of a CoP-aware reader. I further adopted to the practices of the KWARC group, such as the formal representation format OMDOC, which allowed me experience the learning and adapting to a new environment firsthand.

Phase I: LECTORA LIGHT. I intend to spend the next 2 years on the execution of the phases towards LECTORA by means of two educational case studies. The first phase is a preparation phase in which practices of researchers, scholars, and students are formalized and user-specific practices are modeled. The first phase aims at building a light version of the reader that uses the information on individual user practices to provide practice-oriented services based on the GenCS scenario. At the end of the phase, LECTORA LIGHT is evaluated by student volunteers. The first phase will allow me to build up an expertise on mathematical practices and gain intuitions of how LECTORA LIGHT may improve existing approaches, such as the approach followed by ACTIVEMATH. Furthermore, I will be able to evaluate my assumptions on whether mathematical practice can be inscribed into documents and whether our semantic representation format facilitates the automatic detection of practice.

Phase II: LECTORA The second phase aims at improving the reader by modeling COPs based on the GenCS students' profiles. The description of the COPs shall then be used to identify its members. Furthermore, community-aware services are identified and implemented in the improved prototype LECTORA, which is evaluated during another case study. Thus, the second phase is used to verify whether the theory of Communities of Practice can be used to improve the information access by community-aware services such as recommendations, ranking of mathematical search results, and the adaptive presentation of mathematical documents.

Evaluation and Summary. The final 4 months I intend to use for summarizing my findings gained within phase I and II and for concluding the intuitions gained during the prototyping phase into guidelines for the implementation of an interactive reader of mathematical documents. Furthermore, I want to conduct an Added-Value Analysis (cf. section 1.2)in order to contrast the user's benefits provided by LECTORA with the required additional investments of users. The finding of the analysis will then be used to emphasize guidelines for the modeling and application of COPs.

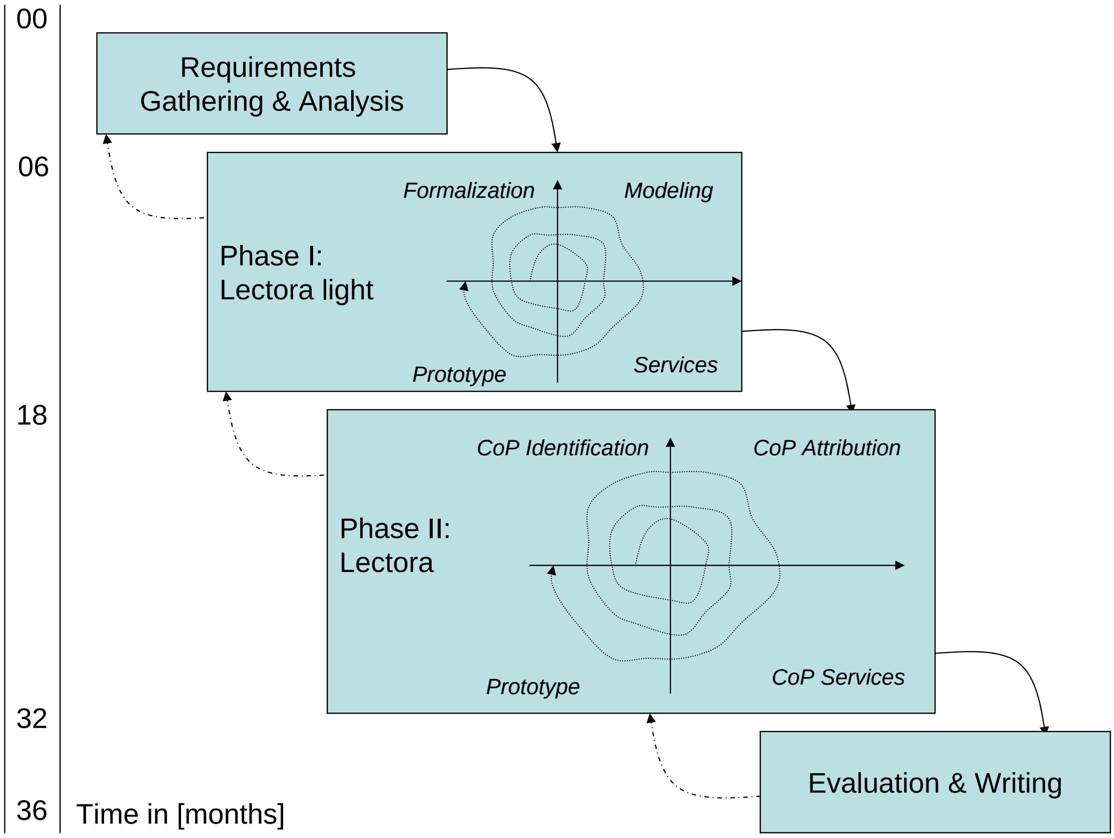

Figure 13: The Vision.

7 References

[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.

[BF06] Jon Borwein and William M. Farmer, editors. *Mathematical Knowledge Management, MKM'06*, number 4108 in LNAI. Springer Verlag, 2006.

[Blo56] B. Bloom, editor. *Taxonomy of Educational Objectives: The Classification Of Educational Goals: Handbook I.* Longmans, Green, New York, Toronto, 1956.

[BP00] D. Billsus and M. Pazzani. User Modeling For Adaptive News Access. *User Modeling and User-Adapted Interaction*, 10:147–180, 2000.

[CNX07] CONNEXIONS. web page at `http://cnx.org`, seen January 2007.

[con07] Connotea, seen March 2007. homepage at `http://www.connotea.org/`.

[DFK07] Deutsche Forschungsinstitut Für Künstliche Intelligenz, seen March 2007. homepage at `http://www.informatik.uni-bremen.de/dfki-sks/`.

[GG99] Harald Gapski and Gernot Gehrke. Media competence. a new concept for framing information society. *European Cultural Foundation (ed.) Newsletter*, Jg. XXII(1), 1999.

[Hei00] Bettina Heintz. *Die Innenwelt der Mathematik. Zur Kultur und Praxis einer beweisenden Disziplin.* Springer-Verlag Wien, 2000.

[HKS06] Eberhard Hilf, Michael Kohlhase, and Heinrich Stamerjohanns. Capturing the content of physics: Systems, observables, and experiments. In Borwein and Farmer [BF06].

[JK05] Klaus Jantke and Rainer Knauf. Didactic Design though Storyboarding: Standard Concepts for Standard Tools. In *Proc. of the 4th Winter Internationals Symposiums of Information and Communication Technologies, First International Workshop on Dissemanitation of E-Learning Technologies and Applications*, pages 20–25, Cape Town, South Africa, 2005. Computer Science Press.

[JRS05] Sabina Jeschke, Thomas Richter, and Ruedi Seiler. Konzepte Intelligenter Assistenten In Virtuellen Laboren für Mathematik. 2005.

[Kit88] Philip Kitcher. *Mathematical Naturalism*, pages 293–325. University of Minnesota Press, 1988.

[KK06] Andrea Kohlhase and Michael Kohlhase. Communities of Practice in MKM: An Extensional Model. In Borwein and Farmer [BF06], pages 179–193.

[KM07] Andrea Kohlhase and Normen Müller. Added-Value: Getting People into Semantic Work Environments. In *Emerging Technologies for Semantic Work Environments: Techniques, Methods, and Applications*. Idea Group, 2007. To appear; chapters under review.

[KMA06] Karsten Krutz, Christian Maier, and Sebastian Abeck. Living documents: Flexibles lernmedium für innovative lernszenarien, 2006. Available at `http://www.cm-tm.uka.de/CM-Web/07.Publikationen/%5BKM+06%5D_Living_Documents_Flexibles_Lernmedium_fuer_innovative_Lernszenarien.pdf`.

[Koh06] Michael Kohlhase. OMDOC – *An open markup format for mathematical documents [Version 1.2].* Number 4180 in LNAI. Springer Verlag, 2006.

[KWA07] Knowledge adaptation and reasoning for content, seen February 2007. homepage at `http://kwarc.eecs.iu-bremen.de/`.

[LMU01] Paul Libbrecht, Erica Melis, and C. Ullrich. Generating Personalized Documents Using a Presentation Planner. In *ED-MEDIA 2001-World Conference on Educational Multimedia, Hypermedia and Telecommunications*, pages 1124–1125, 2001.

[LSY03] G. Linden, B. Smith, and J. York. Amazon.com Recommendations: Item-to-Item Collaborative Filtering. *IEEE Internet Computing*, (1):76–80, 2003.

[LW91] Jean Lave and Etienne Wenger. *Situated Learning: Legitimate Peripheral Participation(Learning in Doing: Social, Cognitive and Computational Perspectives S.).* Cambridge University Press, 1991.

[MAF+01] E. Melis, E. Andres, A. Franke, G. Goguadse, P. Libbrecht, M. Pollet, and C. Ullrich. ACTIVEMATH System Description. In Johanna D. Moore, Carol Luckhard Redfield, and W. Lewis Johnson, editors, *Artificial Intelligence in Education*, volume 68 of *Frontiers in Artificial Intelligence and Applications*, pages 580–582. IOS Press, 2001.

[Man77] Yuri I. Manin. *A Course in Mathematical Logic*. Springer-Verlag, New York, Heidelberg, Berlin, 1977.

[Mat07] W3c math home. Web site at `http://www.w3.org/Math/`, seen March 2007.

[MBA+01] Erica Melis, Jochen Büdenbender, E. Andres, Adrian Frischauf, George Goguadze, Paul Libbrecht, M. Pollet, and C. Ullrich. ActiveMath: A Generic and Adaptive Web-Based Learning Environment. *Artificial Intelligence and Education*, 12(4), 2001.

[MKM07] MEETINGS AND CONFERENCES ON MATHEMATICAL KNOWLEDGE MANAGEMENT. Project homepage at `http://www.mkm-ig.org/meetings/`, seen March 2007.

[MLUM05] S. Manzoor, P. Libbrecht, C. Ullrich, and E. Melis. Authoring Presentation for OpenMath. In Michael Kohlhase, editor, *Mathematical Knowledge Management, MKM'05*, number 3863 in LNAI. Springer Verlag, 2005.

[msSO+05] mc schraefel, Daniel A. Smith, Alisdair Owens, Alistair Russel, Craig Harris, and Max Wilson. The Eveolving mSpace Platform: Leveraging the Semantic Web on the Trail of the Memex. In *ACM Conference on Hypertext and Hypermedia*, 2005.

[Net07] Netflix, seen February 2007. homepage at `http://www.netflixprize.com/`.

[Not07] Notation Selection Tool, seen February 2007. available at `http://www.orcca.on.ca/MathML/NotationSelectionTool/`.

[Pol73] G. Polya. *How to Solve it*. Princeton University Press, 1973.

[SW06] Elena Smirnova and Stephen M. Watt. Notation Selection in Mathematical Computing Environments. In *Proc. Transgressive Computing 2006: A conference in honor of Jean Della Dora (TC 2006)*, Granada, Spain, 2006.

[Vis07] Microsoft Office Visio, seen February 2007. available at `http://office.microsoft.com/de-de/visio/FX100487861031.aspx`.

[W3C07] World Wide Web Consortium, seen February 2007. homepage at `http://www.w3c.org`.

[Wen05] Etienne Wenger. *Communities of Practice: Learning, Meaning, and Identity*. Cambridge University Press, 2005.

[Wik06] Wikipedia, the free encyclopedia. `http://www.wikipedia.org`, 2001–2006.

[XPo07] XPointer, seen March 2007. homepage at `http://www.w3.org/TR/xptr-framework/`.

Index