Personalized Access to Distributed Learning Repositories – PADLR – Final Proposal

Authored by all Participants
Coordinated by Wolfgang Nejdl

March 25, 2001

Abstract

The proposal is structured as a framework + modules document. The framework pages give an executive summary of the project, describe the overall research agenda, and briefly list the individual PADLR modules. Module descriptions written by the subgroups responsible describe the research agenda in more detail. Each module description includes explicit information about how it is integrated into the total framework and describe the interaction with other modules and groups (focusing especially on cross-disciplinary and cross-institutional collaboration). The descriptions further specify deliverables, timelines, testbeds and budget information.

1 Executive Summary

Higher education faces problems of increasing student groups in combination with a shortage of resources for lecturing time. As time for student/teacher-interaction is reduced, students have to perform more self-studies than before. This calls for effective modularized and personalized learning materials, which help students to prepare for courses, support self-studying periods as well as own contributions during their courses, and provide additional background information as well as (project) documentation after their courses.

The driving vision for this project is a distributed "learning web infrastructure", which makes it possible to exchange/author/annotate/organize and personalize/navigate/use/reuse modular learning objects, supporting a variety of courses, disciplines and universities. Each of the PADLR subprojects deals with a specific problem on the way towards this vision. Infrastructure, tools, courselets and archives will be designed/developed in accordance with international standards for modularization and metadata, and will be compatible across the PADLR project. We will specify how courselets are built (both from a technical and from an educational point of view), how they are organized and how they are exchanged and reused.

Several testbeds at our universities will be the test and application area of our infrastructure, tools and courselets, providing a rich source of requirements, feedback and evaluation results for steering our project into the right direction. Due to budget constraints, we could not include all testbeds envisioned in our letter of intent in our final proposal (for example "Heat and Power Technologies" at KTH). We hope we will be able to expand our project budget in the second year to include additional project members from the participating sites.

Due to different budget years, project funding is supposed to start on April 1st, 2001 for German and Swedish modules, and on September 1st for Stanford modules, with possibly some earlier startup funding. The project proposal describes work for about two years, end date for proposed work at all sites ist August 30th, 2003.

2 Project Overview

Identified Problems. Higher education (in traditional study programs as well as in continuing education programs) of today faces problems of increasing and often inhomogeneous student groups in combination with a shortage of resources for lecturing and preparation time. As time for student/teacher-interaction is reduced, students have to perform more self-studies than before.

Beginning students are often frustrated by this situation, finding it hard to spend so much of their study time alone with a textbook to fill gaps in their learning, which cannot be taught within the traditional courses. Advanced students on the other hand, who are more competent as learning consumers, are attending courses with strict course outlines and learning materials they already know or do not need, frustrating their desire to have more control over their learning and direct their learning to better meet their personal needs.

Faculty members and teachers are also frustrated, spending a lot of their time on course preparation (usually without being able to reuse existing materials) and on managing large groups of students with different (disciplinary) backgrounds, different learning preferences and different levels of knowledge. Though a lot of course materials are produced using various authoring tools and environments, they tend to be isolated and to be reused only to a very limited extent.

Proposed Solutions. The relation between time for lectures and the time for self-studies will necessarily have to be changed. The time for student/teacher-interaction is, and will be, limited and the students will have to perform more self-studies than before. By increasing the amount of time that students can control in terms of time and place, the university can adopt to the new situation by increasing flexibility.

If we only have time for a limited number of lectures during a course, we have to ensure that students are as well prepared as possible, even if they come from different backgrounds and with different previous knowledge about the subject. Online learning modules / courselets can support this goal, provided that they are of high quality in terms of factual content, pedagogic embedding, interactivity and accessibility in a broad sense. Specific learning modules can be used during the course, extended and customized from a modularized and distributed pool of resources available to the instructors and students. Additional learning modules can support excellent students and further studies.

It is the goal of our proposal to advance the state of the art in instructional design as well as in technological infrastructure to make such a scenario feasible, beneficial and at last common-place.

Approach and Research Strategy. The PADLR Project Proposal consists of several modules and subprojects, each dealing with a specific problem on the way towards our goal. We will work on necessary infrastructure, tools, courselets and distributed archives (to be designed/developed in accordance with international standards for modularization and metadata), and we will make them compatible across the whole PADLR project. We will specify how courselets are built (both from a technical and from an educational point of view), how they are organized and how they are exchanged and reused, and how distributed content archives can be queried and navigated.

We will use several testbeds at our universities for providing a rich source of requirements, feedback and evaluation results. These will be the test and application areas of our infrastructure, tools and courselets, helping us to steer our project into the right direction.

The PADLR subprojects are grouped into three different modules, including the descriptions of testbeds at the participating sites. The proposal consists of three intertwined modules, with several submodules each:

- The module “Infrastructure and Intelligent Services” includes work on exchange facilities and basic infrastructure, personalized queries and views over distributed learning materials, and automatic extraction of metadata and ontological information.
- The module “Server and Client Side Tools” includes work on modular content archives and video/audio capturing and metadata annotation tools.
- The module “Shared and Personalized Access to Educational Media” includes work on personalized learning sequences, interfaces, and guidance, personalized access to large text archives and personalized and shared mathematics courselets.

3 Module: Infrastructure and Intelligent Services.

3.1 Exchange Facilities / Basic Infrastructure

Working Title. Edutella: An Infrastructure for the Exchange of Educational Media

Contributing Research Groups and PIs. Stanford Infolab (Manning/Decker), Hannover KBS (Nejdl), Stockholm CID/NADA (Naeve)

Problem Description. Every single university usually has already a large pool of educational resources distributed over the institutions. These are under control of the single entities, and it is unlikely that these entities will give up this control. Thus central-server approaches for the distribution of educational media are unlikely to happen. To facilitate exchange of educational media, we propose to develop an exchange network for educational media, based on an approach based on peer-to-peer (P2P) networks. These P2P networks have already been quite successful for exchanging data in heterogeneous environments, and have been brought into focus with services like Napster and Gnutella, providing access to distributed resources like MP3 coded audio data.

However, pure Napster and Gnutella like approaches are not suitable for the exchange of educational media. For example, the metadata in Gnutella is limited to a file name and a path. While this might work for files with titles like ”Madonna - like a Virgin”, it certainly does not work for ”Introduction to Algebra - Lecture 23”.

The educational domain, we thus see, is in need of a much richer metadata markup of resources, a markup that is often highly domain and resource type specific. In order to facilitate interoperability and reusability of educational resources, we need to build a system supporting a wide range of such resources. This places high demands on the interchange protocols and metadata schemata used in such a system, as well as on the overall technical structure.

The “Open Archives Initiative”, which has recently defined a HTTP-based protocol for retrieving metadata information about digital documents from library servers, is an interesting initiative, whose further development might also be of interest to the issues discussed in this proposal. Currently, though, it is server-retrieval protocol only, and just only Dublin Core as its basic schema.

Research Plan and Deliverables. To solve these problems we propose to use a metadata based peer to peer system for educational resources, including Edutella nodes providing different client/server functionalities based on common and interoperable metadata and P2P conventions. Constructing such a system will pose two concrete problems: 1) define the metadata infrastructure and functionalities for Edutella and 2) define the peer-to-peer infrastructure and functionalities. For both problems we have to define the basic (abstract) model and to design a sound implementation.

For defining and implementing the metadata infrastructure and functionalities (1), we have the tackle the following tasks:

1. Metadata and metadata schemata for learning resources. We need to specify a basic model (RDF, RDF(S), [22]) to express metadata, to identify / describe metadata schemata (standard schemata like IEEE-LOM (P1484.12) to describe general aspects of learner resources, MPEG-7 for specific aspects of audio-visual content, and area/subject specific metadata like the ACM classification scheme for computer science topics, etc.) and to describe how additional kinds of metadata schemata can be added (e.g. learner models (IEEE P1484.2), which characterize learners and their knowledge/abilities to enable personalized instruction and allow creating and building personal learner models utilizized throughout their education and work life). Using RDF and RDF(S) will be strategically important in order to design our system to be compliant with the next generation (= semantic) web [12, 8, 28]. Interoperability with existing web standards and metadata schemata are crucial requirements to lower the entry level into the system. This is the fundamental part of defining the metadata information handled by the Edutella system (KBS, CID, IfN).

2. Defining the metadata handling capabilities (API, query language, updates) and identify required functionalities for the access to the metadata. Edutella will bring not only multiple and distibuted content, but multiple and distributed meta-data as well (where several different schemata for the same content will be possible and several authors provide metadata for the same content). Metadata therefore have to be identifiable and reproducible to ensure that different meta-data sets are consistently separable, metadata will be seen as a resource on its own (see also the next point on schema interoperabilitiy). This is the fundamental part of accessing and working with Edutella metadata irrespective of particular storage representations. See also [25, 30] (Infolab, CID, KBS).

3. Define and set up metadata schema bureaus, i.e. servers that store metadata schemata and classifications used in the Edutella network (see also [11]. Enabling access to such explicit descriptions of schemata is crucial in order to allow sharing of metadata schemata within user communities [17]. RDF schemata already support this mechanism, RDF metada explicitly reference machine readable descriptions of the schemata they are instantiated from. These schemata will then be available as resources within the Edutella network. (Infolab)

4. Implement ontology/schema mapping capabilities to be used within the Edutella network in order to allow exchange of metadata based on different, but related metadata schemata. (Infolab)

For the peer-to-peer infrastructure and functionalities (2), the following tasks are important:

1. Define a set of well-designed interfaces to the Edutella system and protocols between Edutella nodes, that allow access to and interchange between Edutella resources. Again we can build upon emerging Web standards (like the SOAP messaging protocol which uses HTTP to carry messages that are formatted with XML and thus provides a lightweight standard object invocation protocol built on Internet standards [7], or the new JXTA open source initiative proposed by SUN). This makes it possible to implement Edutella facilities e.g. as plugin modules to SOAP enabled servers (such as Apache). Even more lightweight basic servers can be imagined, for example

using simple WebDAV over HTTP, or a read-only servers serving metadata as static XML files. Integrating such lightweight server capabilities into arbitrary Edutella nodes leads to a P2P network where every node can offer services to other nodes (possibly with different sophistification). This approach satisfies the very important demand that setting up a basic Edutella server should be a very simple procedure, and that different layers of added functionality can be added step by step later. This is the fundamental part of accessing Edutella services over the Internet/Web (see also [13]. (CID, KBS)

2. Specify a number of supported addressing schemata. We need to address the problem of relocation of resources and servers. Especially in the educational domain, there is a fundamental need for persistent, location-independent resource identifiers. A two step procedure, where the resource identifier is separated from the actual retrieval process would allow easy relocation. Path URNs is such a solution, and IMS is working on other proposals in this direction. We need to define which such protocols that are to be supported in the system, and how the metadata nodes should cooperate in the resolving of such identifiers. (CID)

3. Implement an open-source base library and set of plugin modules with support for the basic metadata model for Edutella nodes and the peer-to-peer networking details between Edutella nodes. These libraries/modules will provide basic access and update functionalities for Edutella nodes, and will also be used in other submodules (like “modular content archives” and “lightweight tools”) (KBS)

4. Implement a set of extended functionalities on top of this base library in order to support different types of extensions (e.g. an ontology inference engine or filtering [31]). Functionalities provided by other submodules like “personalization” can also be implemented as such extension plugin modules plus appropriate metadata/metadata schemata. Important further aspects include loadbalancing and efficient querying in the Edutella network, influencing both protocol design and dynamic reconfiguration of network (KBS, Infolab).

Dissemination, Testbeds and Evaluation Dissemination of results will be done through reports and scientific publications on the different aspects outlined in the research plan. A set of prototype implementations at the participating sites as described above will be available after the first year, which will be refined and extended during the second year based on a evaluation and feedback from these implementations. We will use several specific courses as well as existing intra- and inter-university project cooperations as resources for our requirements analysis and as testbeds for our implementations.

In Germany for example, our testbed context will be the ULI project, a distributed computer science program, where 8 universities and 17 CS professors will work together for 3 years to create a distributed computer science study program as well as course content for these courses (financed with about 3 Mill. Euro for the next three years). KBS is one of these partners (responsible for the area of artificial intelligence), and will use this project as a testbed for the infrastructure and tools developed, both in the requirements, refinement and evaluation phase. In Sweden, our testbeds will consist of the modular content archives for humanities and of personalized and shared mathematical courselets.

Collaboration and Scholarly Exchange. Strong interaction with all modules building tools und functionalities, in order to define common standards, strong interaction with all modules working with testbeds in order to define requirements, and to use evaluations to drive development. Use research visits (2 weeks up to 3 months) (Hannover, Stanford, Stockholm) in order to integrate design and development within this module and with other modules .

Budget Overview (including overhead costs):

KBS: 70K first year, 40K second year. Budget will pay for one / part of one Ph.D. Student, L3S infrastructure costs, travel and exchange.

IfN: 20K first year. Budget will pay for part of a Ph.D. student, L3S infrastructure costs, travel and exchange.

Infolab: 70K first year. Budget will pay for one Postdoc, overhead costs, travel and exchange.

CID: 30K first year, 30K second year. Budget will pay for part time Ph.D. Student, overhead costs, travel and exchange.

3.2 Personalized Queries and Views over Distributed Learning Materials

The module will provide an infrastructure for advanced personalized query facilities on top of the Edutella meta-data and peer-to-peer infrastructure. This would complement the deliverables in the module ’Infrastructure and Intelligent Services’.

Working Title. PSELO: A personalized search engine for learning objects.

Contributing Research Groups and PIs. Uppsala University (Risch)

Problem Description. In a distributed learning environment there will be large numbers of learning objects and courselets stored in many distributed and differing data stores on the Internet. Without guidance, students will have great difficulties to find the learning objects relevant for a particular learning task. The meta-data descriptions provide information about properties of the learning objects, but the meta-data by itself does not provide for handling reconciliation of differences between different objects nor does it provide for customized and efficient queries over views of reconciled learning objects.

Research Plan and Deliverables. The purpose of this module is to provide technology for defining personalized learning views of relevant learning objects / courselets for each subject, student, or task. These personalized views focus the data primarily seen by a user to a particular set of relevant objects. The student can then explore the relevant learning objects through a powerful and *subject-oriented* query language. Here subject-orientation means that the operators used in queries are specialized for a particular learning subject. Such a query language provides a tool for the student for advanced exploration of a subject. It should be possible to dynamically change and adapt the personalized view for each student, task etc., as new knowledge is increasingly deeper explored.

The personalized views are defined from the meta-data model. In our approach these views are defined for the user as a set of object-oriented (OO) data views inferred from meta-data. Subject-oriented queries are then specified in terms of the personalized data views using an extensible OO query language.

For defining and implementing such a query infrastructure the following tasks need to be solved:

1. The subject-oriented queries should be expressed in terms of the meta-data model of learning objects. The system must therefore have the ability to interpret meta-data definitions of Eduella. The learning object often have complex structures that need to be viewed as object structures. The query language must thus have the ability to express queries and views over complex object structures.

2. To support subject-oriented queries it must be possible to easily extend the query engine with knowledge modules implementing the subject-oriented query operators. This includes plug-in facilities for implementing both query operators as well as optimization rules for subject-oriented queries.

3. Since the learning objects are stored in many different places and in many different formats the query engine must be able to deal with queries that span many heterogeneous and distributed data sources.

   A major challenge is here to efficiently process subject-oriented queries that access and process many distributed and heterogeneous knowledge objects. The query execution should thereby utilize the peer-to-peer infrastructure of Eduella.

The project enhances data integration technology for support of learning applications and is expected to enhance the state-of-the-art in the areas of database and data integration techniques. It will produce reports and scientific publications.

The work will extend on research developed at Uppsala Database Laboratory (http://www.dis.uu.se/˜udbl) on OO queries over distributed and heterogeneous data [41, 42, 32]. Research from other institutes on wrapping and searching heterogeneous, distributed, and semistructured data, e.g. [15, 23, 40], is also applicable. Other important supplemental components are user interface modules for subject-oriented visual query specifications and for visualizing retrieved learning objects, which can be added as plugins.

Dissemination, Testbeds and Evaluation In the project we will implement a prototype search engine, PSELO, that provides personalized OO views of the distributed learning objects. The prototype system will demonstrate the feasibility of the approach and serve as a platform for further experimentation and evaluations.

It is envisioned that this technology will be an important component of the Edutella infrastructure. It would utilize the meta-data protocols of Edutella for providing the terminology in which the personalized views and queries are expressed. It will use the Edutella peer-to-peer infrastructure for efficient access to learning objects from the queries. With the proposed system it will furthermore be possible to incorporate algorithms for a particular subject as system plug-ins that define subject-oriented query operators.

Collaboration and Scholarly Exchange. The work on this module is closely related to other modules building tools and functionalities. Research visits to other participants are expected to be complemented with dayly e-mail contacts.

Budget Overview (including overhead costs):

Uppsala: 35K first year, 35K second year. Budget will pay for half a Ph.D. Student, infrastructure costs, travel and exchange.

3.3 Automatic extraction of metadata and ontological information

Contributor. Chris Manning (Infolab/CS, Stanford)

Working Title. Extract: Automatic extraction of metadata and ontological information.

Problem description. The basic infrastructure portion of the proposal depends on the availability of accurate metadata for learning resources, built to fit within a loose ontology. While a metadata framework provides a lightweight solution to many of the problems to be addressed in this project, it leaves the question of where the metadata is to come from. For core content, hand-annotation is possible. However, even for core content, this will be difficult: busy educators will not appreciate having to carry out an additional task beyond content creation. A reason why many CBL projects fail is because the materials change more slowly than educational needs. *Any tool that can reduce this burden by semi-automatically providing metadata will be useful*. The problem is even more acute for the enormous world of information (e.g., on the Web) outside the core resources. There is already a vast array of useful teaching resources on the web, and students could often get value from making use of advanced or complementary materials at other institutions. However, connected to the personalization issues of adaptation comes the difficult and very time-consuming tasks of finding appropriate materials, and determining their prerequisites, etc. In this context, standard keyword search is of very limited effectiveness, because it cannot filter for: (i) the *type* of information (tutorial, applet or demo, review questions, etc.), (ii) the *level* of the information (aimed at secondary school students, the general public, or graduate students?), (iii) the *prerequisites* for understanding the information, or (iv) the *quality* of the information. Moreover, there are all the usual problems of keyword-based information retrieval, such as problems with synonymy, polysemy, and so on. *Any method which automatically annotates information from other sources so that it can be easily accessed within our content repository will be of enormous value*. Such additional resources, while of less consistent overall quality, will magnify the value of our content repository.

Research plan and deliverables. The starting point is the use of statistical information extraction and natural language parsing techniques to automatically derive classifcatory and metadata information from primarily textual data (web pages, Word, postscript or similar documents, etc.). While still challenging for large ontologies, text classification methods which semantically categorize an entire document are now relatively well-understood, and provide a good level of performance. A central research challenge is how to extend these methods to address issues like trying to determine the prerequisites for understanding a topic, or the quality of the information presented. We believe that two of the most important sources of information here will be linkage information (in the case of hypertext sources like the Web), and natural language processing terminology extraction (to identify terms that are assumed without definition within the text). Determining the level of information is also little explored, beyond crude measures of reading level. Exploring NLP and machine learning approaches to such problems will be the main focus of the research.

The research will lead to a prototype system which will classify educational resources from outside the repository with the metadata attributes defined in the central Edutella infrastructure.

Dissemination and Evaluation. The results of the research will be disseminated through scientific publications. The prototype will be integrated with the Edutella infrastructure in the second year, and evaluated for its success in finding and classifying suitable information. In the first instance, this success will be measured using information retrieval evaluation methods. A more task-specific evaluation would involve demonstrating the additional value over existing resources provided by returned results, as seen in user studies. This evaluation can be an aspect of the more general evaluation of the testbeds.

Collaborations. On the one hand, this module fits in strongly with aspects of the Edutella infrastructure, in particular the issue of ontology mapping (InfoLab), which requires similar techniques. On the other hand, resource identification and accurate classification, in particular identifying approaches and prerequisites are important enabling technologies for effective personalization, and this project will work closely with KBS and Uppsala groups on personalized content delivery and computational linguistics methods. Research visits will assist in integrating the design and development between these two groups.

Budget Overview (including overhead costs):

InfoLab: 70K second year. Budget will pay for one PhD student, overhead costs, and travel.

4 Module: Server and Client Side Tools

4.1 Modular Content Archives

Contributing Research Groups and PIs. IfN Braunschweig (Ulrich Reimers), KBS Hannover (Wolfgang Nejdl), Uppsala (Broady), CID (Broady)

Working Title. Archives: Intuitive Storage, Use and Retrieval of Archived Educational Media.

Problem Description. At every university usually plenty of material is available or being produced for the planning, preparation and execution of curricular activities like lectures, seminars and project work. This material is an inhomogeneous amount of content of various type. These could be lecture slides, presentations, scientific graphics and texts, description of experiments, software, animations, simulations.

Existing distributed learning environments often demand content to be molded into proprietary and application dependant formats. Therefore it is difficult to re-use content for different purposes and different audiences. Hence, teachers today are prone to produce archives and curricula based on redundant information, proprietary applications and formats, and non-modularized solutions.

An intelligent and flexible archiving, management, allocation and distribution of this modular content is an intense problem. Several commercial products like Hyperwave or Lotus offer certain functionality providing a step into the right direction but there is still the lack of a sophisticated archive system to optimally support curricular activities. As mentioned earlier in this proposal a central server approach is due to the distributed nature not the optimum type of architecture, and - building on the Edutella infrastructure - one of the distinguishing feature to other projects (e.g. the German “Teachware-on-Demand” project) is exactly this distributed peer-to-peer infrastructure, with plugins for different kinds of peers participating within the Edutella network.

Intelligent archiving and flexible and intuitive access to modularized content come along with several problems: Content modules are often detached from context so that interrelated content modules can hardly be found and are therefore of no use. This for instance will prevent an instructor to answer a students question with appropriate instructive material if this material is not intended for that particular course. Content modules can be of different file types. Locally distributed creation leads to the need for version lists. Offline archiving of created content modules is difficult and tedious. Instructors have no fast access to content modules due to the lack of appropriate search strategies and missing context information. Students have no access to many interesting content modules. Especially colleagues from Sweden and Germany will produce same content in different languages that should be linked.

Another problem is that in many educational settings teachers and students are not able to profit from the international development of agreed markup schemes that are evolving within research communities (mainly SGML-based or XML-based, such as the TEI encoding guidelines among scholars in the humanities). Since the tools and practices used in courses often do not keep pace with such international de facto standards, the well-structured content of existing and emerging digital research archives are in many cases not easily available to teachers and students.

Yet another problem concerns the rapidly increasing volume of scientific results/deliverables available at the WWW. In some rapidly evolving fields, such as bioinformatics, there are hardly any course books. This means that curricular development will be dependent on teachers' ability to overview, navigate, identify and re-use appropriate selections of existing digital archives. There is a need for better principles for the design of modularized content repositories.

There are also a series of more narrow technical problems. Since existing materials are often stored in various file formats and developer versions along with non-valid filenames, it is difficult and sometimes impossible to re-use information.

Research Plan and Deliverables To solve the described problems the following tasks have to be accomplished:

Metadata and Windows Applications. A strong focus lies on the integration of (Windows) de-facto standard file formats like Word, PowerPoint, JPG, GIF, HTML, Mathcad, Windows executables etc. into Edutella archives. We will investigate, which kind of constraints this leads to, how Edutella metadata are used especially in this environment.

To access this content very fast efficient and flexible text/keyword search facilities are necessary. One situation that occurs during a lecture is the student having a question that deals with topic outside the primary scope of the lecture. Using these sophisticated search facilities instructors can access appropriate content very fast and answer the students question instructively. Another important aspect is context information for content modules. If every single module can have appended context information, it will be possible to provide fast access to connected modules via visual presentations (context maps).

The described tasks will lead to intensive cooperation with the ”Edutella” and the ”Personalized Interfaces” modules. (IfN)

Repository Architecture and Tools. We propose a local database system with a client server architecture to support instructors in an optimum way. This single database functions as a university wide repository that can exchange metadata and content with other distributed repositories as described in the ”Edutella” module. Commercial products like Hyperwave will be examined if these can be extended to suit our needs. XML/SOAP et al can function as a standardized protocol for the exchange of metadata and data between repositories or a repository and a user. Similar developments at other institutions will also be taken into account (for previous work see e.g. [27]).

An important issue is the linking of personalized electronic student portfolios with such a repository. Students will be able to download interesting material to their own electronic portfolio and manage their content. These issues will be discussed in cooperation with members of the METAFOLIO group. (IfN)

Another task is to develop and explore methods and tools for the design and creation of modularized content archives, as well as develop and explore methods and tools that allow teachers and students to access and use already existing repositories on the web, with a strong focus on modularity and re-usability in different context, based on the content archive systems described in the previous paragraphs. They should support the teachers and students to work with these archives. For a specific course the teacher might propose the students certain paths through the archive and certain subsets of content modules to be used by the students and in some cases added to their portfolios. (IfN, KBS, CID, Uppsala)

The knowledge management tool Conzilla already supports the organization of annotated content into personalized portfolios in a way that is compliant with the emerging standards for automated information exchange (XML) and metadata handling (IMS). We will modify Conzilla to enhance this support in various ways, based on the structure of this repository, and on user-feedback from our activities in other modules. (CID)

Metadata-Based Learning Repositories. Based on the Edutella submodule we will implement a second server implementation called an open learning repository (OLR), which just stores (RDF) metadata (both for classification/annotation and for structure) in a central database, but no content (which is accessible via URLs/HTTP).

Resources potentially distributed all over the Internet can be combined to entities we call "Courses", logical documents integrating content from different sources. While the content stays at its original location our database holds all information about the structure of each course and metadata about the content. The responsibility for maintaining content stays with the people providing it. When browsing a course all HTML-pages are generated dynamicall.

The system provides a platform for testing out different navigation schemes including traditional concepts such as hierarchical tree structures and tutorial-like course-trails as well as highly adaptive approaches (personalized navigation) – see the personalized access submodules. (KBS)

A second system will explore an alternative architecture based on XML, XML Schema and JSP, again compatible with the Edutella functionalities. It will include a WebDAV-based courseware authoring module which enables geographically dispersed courseware authors to collaboratively work on the course contents, and a standalone, JSP-based courseware publishing engine which can achieve flexible, dynamic courseware presentation. The system can be aware of any changes of the course contents and automatically reflect the changes on the Web at any time, again customizable by metadata. (KBS, CID)

Dissemination, Testbeds and Evaluation Dissemination of the achieved results will be accomplished by scientific publications in appropriate journals and by demonstrations at symposia.

In several testbeds we will show the impact of the developments in this module on teaching and learning. In Braunschweig, we will use several courses in the field of communications, encouraging exchange to other interested faculties from other universities, in Hannover we will use several computer science courses and the ULI project. At Uppsala University several courses in humanties and social sciences and bioinformatics will provide suitable testbeds.

Collaboration and Scholarly Exchange Interactions with other modules:

1. Edutella module (Exchange Facilities / Basic Infrastructure)
2. Automatic extraction of metadata and ontological information
3. Personal Search Engine

There are also close links between student portfolios and content archives. The students should be offered (and themselves able to create) content in portable modularized formats, suitable to be incorporated into their portfolios and to be re-used for various purposes that may not be foreseen by the teacher. Therefore the development of portfolio practices presupposes the creation of archives of content modules.

Use research visits (2 weeks up to 3 months) (Braunschweig, Hannover, Stanford, Stockholm, Uppsala) in order to integrate design and development within this module and with other modules. Workshops will be suitable to promote collaboration as well.

Budget Overview (including overhead costs):

IfN: 50K first year, 70K second year. Budget will pay for one Ph.D. assistant (part-time first year), L3S infrastructure costs, travel and exchange.

KBS: 20K first year, 40K second year. Budget will pay for a part-time Ph.D. student, L3S infrastructure costs, travel and exchange.

Uppsala: 15K first year, 15K second year. Budget will pay for a part-time Ph.D. student, infrastructure costs, travel and exchange.

CID: 15K first year, 15K second year. Budget will pay for a part-time Ph.D. student, infrastructure costs, travel and exchange.

4.2 Video/Audio Capturing and Metadata

Working Title. VACE: Video/Audio Capturing and Embedding

Contributing Research Groups and PIs. Hannover RVS (Pralle)

Problem Description. Currently we see the need to integrate captured audiovisual material from actual seminars, courses, lectures etc. into on-line (web-based) and off-line (e. g. CD-ROM-based) learning systems.

These A/V recordings can help students in understanding the content, especially when they are embedded in other kinds of learning material, as a simple example, the script of the course. In addition this, a larger group of learners can be reached and students can get access to previous material.

But nevertheless, it does not help to just do a recording of a course an publish this on the web, a CD-ROM or a video tape. The key for a useful solution is the integration in a multi-media learning system. The idea is not to produce a “TV-Show” but to use captured media as a supplement to other learning material.

There are some approaches that all have several advantages and disadvantages.

1. One way is to put all information (the A/V-recording of the lecture and the slides) in a TV-quality audio/video stream, like it is done in a project called “Uni-TV” (http://www.lrz-muenchen.de/wir/gigabit-vpp/demo-unitv/).

   There are certain drawbacks in this “single-media” approach: The resolution of the slides is restricted to the TV standard (about 720 576 pixels). A high bandwidth is needed to deliver the media stream (at least four Megabits per second for a MPEG-2 stream). Apart from a VCR-like control (pause, play, fast forward, rewind) the user cannot navigate the content. Finally, there is no association to other media types possible.

2. There are certain presentation tools commercially available (Real Presenter, Microsoft Media Tools) that allow to record an actual presentation and generate an on-line representation of this. Two media streams are produced, one containing the slides, the other one containing the A/V recording of the instructor.

   As the produced media streams are vendor-dependent, they cannot be easily integrated in own learning applications. The re-use of the generated material is difficult or even not possible because it the produced data depends on the format used by the tool. The quality of the streams is often poor because the tools are mainly developed for low-bandwidth transport.

3. One general approach is to record the presentation on video, convert this to a format suitable for online delivery and link it to web pages containing the slides of the presentation.

   Especially if different media types are used much effort has to be spent on the post-production. The slides (e.g. in MS Powerpoint format) have to be converted to browseable web pages and the recorded audio and video data has to be encoded to different streaming formats.

4. There is some development going on in building a system for the “authoring-on-the-fly” of lectures. Several media streams are captured (audio, video, whiteboard) an can be published right after the presentation.

   This LINUX-based systems needs special software on both, the instructors and the students computers. Common applications, file- and streaming-formats cannot be used with this system. Although a platform-independence is not urgently necessary, the applications needed to obtain the media should at least run on common PCs in order to achieve a wide acceptance.

As we see, although it seems to be a common problem there is no ideal solution yet.

Research Plan and Deliverables. To come closer to a solution the features of ideal capturing systems has be seen from different kinds of view.

1. The authors (i. e. the instructors or professors) view. In the terms of ”light-weight-authoring” the system has to be easy to use for the producer. In an ideal case, the capturing tool should run in the background so that the professor can focus on the actual presentation. Common presentation tools like MS PowerPoint, PDF viewers or web browsers should be supported and associations to the captured A/V content should be generated automatically. The post-production should be reduced to a minimum. The content should be reusable for different kinds of media (e. g. on-line/web, off-line/CD).

2. The users (students) view. The viewer should have the possibility to navigate and browse the content. Different ways of using the content have to be offered to the learner to satisfy the individual preferences: Watching the course like a video (“timeline-based”) or reading, browsing and searching it like a book. The slides and the A/V streams should be synchronized to each other so that the viewer can choose between watching the AV stream or browsing through the web pages without losing access to the associated other data type. The integration of other systems like discussion-forums or chat could be helpful.

In order to construct a flexible, open system it would be useful to think of several modules with defined interfaces. There has to be modules for capturing meta-data, interfaces to A/V-systems and publishing. The different modules make use of different existing applications for presenting, A/V-codecs and streaming-systems but itself will be vendor-independent.

Example scenario: The author uses PowerPoint to present his slides in the actual presentation. The lecture is recorded by a PC with capture card. A browseable web presentation and a CD-ROM with additional learning material should be created.

In this case we need a module that captures the events in PowerPoint (changing slides) and extracts the content of the slides for web pages. Another module has to generate meta-data for A/V streaming (e. g. RealMedia) by associating the time-stamps and keywords from the slides. A third module generates an HTML framework for the web presentation that contains the content from the slides and the links to the A/V streaming media. Finally a module that extracts the information for the CD-ROM production is needed.

The concept of modules also allows the integration of forthcoming techniques for media delivery. Metadata will be recorded in a way compatible to the Edutella submodule.

The generated content should be accessible by the use of common applications like web-browsers an streaming players like Real or Windows-Media so that no special programs have to be installed on the students computer.

The content can be use for different purposes, e. g. for archives, portfolios and can be integrated in existing or forthcoming learning systems. The system can be seen as a framework. Extensions as chat, forums or other communication systems can be integrated when using on-line publishing.

One could think of on-demand delivery in web-based on-line systems as well as the use in a real-time learning scenarios or for stored media (on CD, DVD).

Collaboration / Scholarly Exchange. Collaboration especially with the Edutella module, based on direct collaboration and exchange visits to other participating labs. Use of this tool in the different testbeds (e.g. ULI).

Budget Overview (incl. overhead costs):

RVS: 70K for first and second year, which pays for one full Ph.D. student, including L3S overhead costs, travel and exchange.

5 Module: Shared and Personalized Access to Educational Media

5.1 Personalized Learning Sequences

Working Title. PLeaSe: Personalized Learning Sequences

Contributing Research Groups and PIs. Hannover KBS (Nejdl/Henze), Uppsala Linguistics (Borin)

Problem Description. It is generally agreed that it is a desirable goal in any educational setting to be able to tailor courses and course materials to individual students' needs, as determined by such factors as their previous knowledge of the subject, their learning style, their general attitude, as well as their cultural and linguistic background. A skilled teacher will be able to achieve this goal in one-to-one interactions with learners, but not as a rule in the lecture hall/classroom settings more typical of higher education. Nor will a developer of traditional course materials be able to cater for individual learner needs to any great extent.

ICT-based educational materials can potentially be much more flexible than traditional course materials, however. They offer a unique opportunity to achieve the mentioned goal through *personalization*, where both the selection and presentation sequence of the units of educational material making up a courselet or course are determined by a *dynamically updated learner model*, which takes into account at least the learner's previous knowledge and her progress in mastering the material.

The aim of the personalization module is to provide individualized access to learning materials, to courses, courselets, etc. A "one-size-fits-all" learning path through these materials (or even through parts of it) would neglect individual needs, knowledge or preferences of the users [21]. To maximize the students' success with the open learning repository it is necessary to provide a quick, user-focussed access to those entities in the learning repository which correspond to the user's actual information needs, to her/his knowledge, current situation and preferences.

Research Plan. Research in adapting information systems to the individual users is conducted e.g. in the area of adaptive hypermedia systems. Adaptive hypermedia systems use a model of the user to collect information about her/his knowledge, goals, experience, etc., in order to adapt the content and the navigational structure. There are two main adaptation strategies in adapting hypermedia systems [9]: link-level adaptation calculates useful navigational strategies for the individual user, content-level adaptation individually places content-chunks on the information entities. We assume that our repositories will provide lots of information on any one topic: different explanations, examples, use-cases, etc., therefore in PLeaSe module we will concentrate on selecting and presenting the most appropriate material for each user into personalized learning sequences [10]. In module PALaTe we will explore the issue of creating new hypertext pages out of existing text, particularly in connection with the large text archives testbed [35].

The overall problem setting for our adaptation functionalities is the following: We will have access to a large set of learning materials. Many different authors can modify, delete or add new content to the OLR, the learning objects might be distributed, and we can expect that we will often find more than one learning object on the same topic.

Adaptive hypermedia systems normally combine learning materials with reading sequences, didactical rules, pre- and post-knowledge or pre- and post-learning objects [19]. As we want to adapt materials from different courses and from different contexts, we need a flexible approach to adaptation [18, 20], where information about learning objects is read out from metadata, but also inferred from the objects themselves. Using information retrieval, information extraction and natural language processing methods for obtaining automatic or semi-automatic indexing is of advantage (see the submodule "Automatic extraction of metadata and ontological information"), even crucial in the case of large textual resources as used e.g. in the Languages/Humanities (see the section on the "PALaTe submodule/testbed").

Deliverables.

1. *User orientation:* The adaptation component will focus on orientation guides: Provide access to relevant information, select and show case studies, give hints, show examples, show context, generate reading sequences.

   - How to deal with different teaching strategies in one single course? Different authors of courses/courselets will follow different teaching strategies.
   - How to deal with information oversupply? Present "best", find best fitting tour.
   - How to generate personalized learning sequences based on different learning theories (see module PerINaG)?
   - How to deal with structured materials? In our repositories we will find lots of structured materials. How can we make use of these structures (for presentation, visualization, selection of materials?).

2. *User Interactions:* All interactions of a user with our learning repositories can possibly carry information about the user's goals, knowledge, preferences.

   - How to interpret this *raw* input data, e.g. mouse clicks?
   - We will use different kinds of user interaction as evidence (and therefore as input for the user model): quiz answers, project performance, retrieved learning objects. We want to distinguish the different strength of belief in conclusions about user interactions. This implies a requirement for the knowledge model: The knowledge model must be able to deal with observations of different kinds and different granularity.

- Diagnostic testing methods (such as the Didax system being developed in the Swedish Learning Lab (SweLL) APE-DRHum project [5, 6]) will be applied to enhance the value of information we obtain from the different user interactions.

3. *User Model / Knowledge Model:* The user model stores individual data like name, overall preferences, abilities, etc. The knowledge model manages learning dependencies of some application domain. In defining these models and implementing tools for using them, we will build upon work done by IEEE LTSC Working and Study Groups, especially P1484.2 (Learner Model WG), P1484.4 (Task Model WG), P1484.6 (Course Sequencing WG), and P1484.20 (Competency Definitions SG).

    - How can we build knowledge models that can be enlarged with new topics?
    - How can we build knowledge models and exchange information with different or partly overlapping knowledge models?
    - Can we construct our knowledge models from the ontologies described in the basic infrastructure module?

Dissemination, Testbeds and Evaluation Dissemination of results will be done through reports and scientific publications on the different aspects outlined in the research plan. A set of prototype implementations at the participating sites as described above will be available after the first year, which will be refined and extended during the second year based on a evaluation and feedback from these implementations. We will use several specific courses as well as existing intra- and inter-university project cooperations as resources for our requirements analysis and as testbeds for our implementations.

In Germany our testbed context will be the ULI project (see description of submodule "Exchange facilities / Basic infrastructure").

In Sweden, Languages and Humanities education will serve as an important testbed, with personalized access to large text archives as the common motif (see the section on the "PALaTe submodule/testbed").

Collaboration and Scholarly Exchange. Interaction with the Module "Infrastructure and Intelligent Services": incorporating requirements from the adaptation component: content and usage (including context of use) of each learning object; indexing (semi-automatic), information retrieval methods in the meta-data definition. In addition, users should be enabled to store and manipulate retrieved learning materials. These learning materials might be structured (hierarchical, sequential, concept maps, etc.).

Interaction with the Module "Interfaces and Navigation": Outcome of the adaptation component (reading sequences, access to examples, alternative views, explanations) - how to integrate it in a smart user interface? Interaction with "Personalized Access to Large Text Archive"s: How can textual resources be effectively used for personalized learning? Use research visits (2 weeks up to 3 months) (Hannover, Stanford, Uppsala) in order to integrate design and development within this module and with other modules.

Budget Overview (including overhead costs):

KBS: 20K first year, 30K second year. Budget will pay for one part-time Ph.D. student, L3S infrastructure costs, travel and exchange.

L3S central: 20K first year, 20 K second year, for a part-time Ph.D. student in media and design education, L3S infrastructure costs, travel and exchange.

Uppsala: 30K first year, 30K second year, budget will pay for a half time Postdoc, overhead costs, travel and exchange.

5.2 Personalized Access to Large Text Archives (Submodule/Testbed)

Working Title. PALaTe: Personalized Access to Large Text Archives

Contributing Research Groups and PIs. Uppsala (Borin/Broady), CID (Broady).

Problem Description. Text is still important in the teaching of almost any subject, viz. in the form of textbooks and other course texts. In Languages and Humanities education, (large) textual resources are also quite often objects of study in themselves. Arguably, their effective deployment as study objects in the context of ICT-based personalized learning demands some kind of language understanding. Hence, personalized access and navigation among such resources should – almost by definition – make use of Computational Linguistics (CL) / Natural Language Processing (NLP) techniques, to complement the more general personalization tools which will be developed in the submodule “PLeaSe: Personalized Learning Sequences”.

In this submodule/testbed, we thus consider the issue of *personalized access to large text archives* in Languages and Humanities education. In order to make the fruits of our labor in the proposed project useable also in other subject areas, we will focus on certain aspects of this issue, namely how (aspects of the) content and difficulty of texts or parts of texts can be inferred and utilized for creating personalized access to text material.

Research plan and deliverables. We will consider the use of two fairly different kinds of large text archives:

1. In language education and linguistics, large text archives are important mainly (but not only!) because of their (linguistic) *form*. Here, the so-called *text corpus* has become an important educational (and research) resource. The uses of text corpora in language education are manifold:

   - as a data source for the preparation of (monolingual or bilingual) word lists, grammars [1, 2, 3], test items (e.g. for diagnostic tests such as the Didax system being developed in the Swedish Learning Lab (SweLL) APE-DRHum project [5, 6]), etc.
   - as a source of empirical examples in ‘data-driven learning’ [4]. The English Department at Uppsala uses the British National Corpus in this way, and other language departments are getting ready to do the same, e.g. the Slavic Department for use in their Russian courses.
   - as a source of reading matter, user-adapted as to its level of difficulty and subject area (where content obviously becomes important, too)

2. On the other hand, in such Humanities subjects as History, Literature Studies, History of Science, Teacher Training, etc., large text archives are important mainly because of their *content*, i.e. because of the information contained in the texts (and, as a rule, the range of languages dealt with will be much smaller; see below). Typical issues which arise when such text archives are to be used in education (or research) are:

   - Locating texts or text portions in the archive which deal with a particular person, place or time (‘PPT extraction’). Partly, this is addressed in the field of Information Extraction (under the heading of “name recognition”), but the problem is still a long way from being solved, especially if we take into account—as our ambition should be—the general problem of entity references in text (by noun, pronoun, hyperonymy / hyponymy, etc.). Concretely, this has been an issue both in the work on the electronic version of Swedish author August Strindberg’s collected works at KTH in Stockholm, and in the work with the so-called Wallenberg Interviews (interviews with Jews who escaped from Hungary and Nazi persecution thanks to Raoul Wallenberg) in the History Department at Uppsala. These and other large textual resources would see more use in education—bridging the gap between education and the kind of research for which this education is preparing the students—if the access to the resources could be made less unwieldy.

- Selecting texts or portions of texts in the archive which deal with a particular topic, or succession of topics, the latter for assembling a reading sequence out of a larger textual material.

For both kinds of text archives, and for many of the issues just listed, methods and tools from the fields of Information Retrieval, Information Extraction and CL/NLP are available. There are also more open-ended research issues in the list, e.g. the—already mentioned—problem of entity references in text, or that of determining the level of difficulty of a text (for a language learner having a particular linguistic background; see also submodule "Automatic extraction of metadata and ontological information", where the related issue of "determining the level of information" is discussed).

Generally, we believe that the realistic course of action here is to pursue so-called 'shallow', or 'knowledge-light' techniques for text corpora used in language education, because of their potential application to a large number of languages—Uppsala University currently offers courses at various levels in about 40 languages—which in practice precludes the use of 'deep', 'knowledge-intensive' techniques. When there are such techniques available (as may be the case for English, German and a few other languages), they should be considered, of course, but developing them from scratch is too costly. For the case of general Humanities textual resources, however, we should consider developing more knowledge-intensive methods for selected problems, such as the 'PPT extraction' already mentioned, where there is an expressed need among educators and researchers.

The work with large text archives will proceed along two interconnected lines of research:

1. We will explore the issue of using partial parsing and information extraction techniques for marking text portions for persons, places, and times, and carry out formative evaluation of these techniques in an educational setting. This work will be pursued in collaboration with the work in the submodules "Automatic extraction of metadata and ontological information" and "PLeaSe: Personalized learning sequences".

   Deliverables: Prototype person/place/time partial parser ('PPT extractor'), and evaluation reports.

2. We will pursue the issue of how to (operationally) define and determine the level of difficulty (or "level of information"; see above) of a text or a portion of a text (for language education purposes it would be useful to be able to determine this even for small linguistic units such as phrases or clauses), and carry out formative evaluation of this definition in an educational setting. This work, too, will be a collaboration with the work in the submodules "Automatic extraction of metadata and ontological information" and "PLeaSe: Personalized learning sequences".

   Deliverables: Preliminary operational definition of level of difficulty (for particular foreign/second language learner), prototype application for determining level of difficulty at least for Swedish and English text material, and evaluation reports.

Dissemination, Testbeds and Evaluation Dissemination of results will be done through reports and scientific publications on the different aspects outlined in the research plan. In general, we plan to do research/development and evaluation in parallel (i.e., formative evaluation), but for obvious reasons, the first year will be devoted mainly to research and development, while the second year will be dominated by deployment and evaluation in regular education.

We will use existing courses in the departments of the Faculty of Languages, in the History Department and in the Department of Teacher Education as resources for our requirements analysis and as testbeds for our implementations.

Collaboration and Scholarly Exchange. Strong interactions with the submodules "PLeaSe: Personalized Learning Sequences", "Automatic extraction of metadata and ontological information" and "Content Archives".

Budget Overview (including overhead costs):

Uppsala: 25K first year, 25K second year. Budget will pay for one part-time Postdoc, and for faculty involvement in testbed integration in regular Languages/Humanities curricula, overhead costs, travel and exchange.

CID: 10K first year, 10K second year. Budget will pay for a part-time Ph.D. student, overhead costs, travel and exchange.

5.3 Personalized and Shared Mathematics Courselets

Working Title. Personalized and Shared Mathematics Courselets.

Contributing research groups and PIs. DSV at KTH Stockholm (Jansson), CID at KTH Stockholm (Naeve).

Problem Description. This submodule attacks two major difficulties for teachers and learners: the difficulty to share and reuse learning material among students and teachers and across geographical and organizational boundaries and the difficulty to personalize and adapt existing learning material to a particular learning situation. Adaptation is relevant both with respect to the students characteristics and the context of learning Our particular concern is mathematics education. By courselets we mean fragments of courses composed from multimedia explanation modules or content modules in electronic form.

The work will extend on research at DSV on personalized support for learning conceptual modeling [38, 33, 39] and research at CID, where the idea of a concept browser has been developed by Naeve and his team over the past 3 years [24, 25]. A first prototype of a concept browser, called Conzilla, has already been implemented [30, 29]. While conforming to evolving international programming and e-learning standards (such as XML and IMS), Conzilla combines conceptual modeling with annotated access to multi-media based archives in a novel way. This makes Conzilla a powerful basic platform for the kind of problems that our project aims to address. Moreover, since Conzilla is being developed as an open source project, it has the potential to evolve into a widely used tool that provides support for students and teachers in handling multimedia-based archives of digital information. Other related systems developed at CID include PDB (Projective Drawing Board) and CyberMath.

In cooperation with the Advanced Media Technology (AMT) laboratory at KTH, CID is presently coordinating an effort to introduce new teaching methodology into mathematics courses at both the university and the school level (see http://www.amt.kth.se/projekt/matemagi.html).

Research Plan. We articulate seven main concerns for this research: the commitment to shared standards, languages and tools to make sharing and reuse possible on a technical level (relations to the Module on Infrastructure and intelligent services), the management of incrementally growing multimedia content archives, in particular version handling and the handling of problems of structure and navigation (relations to the module on Infrastructures and intelligent services), the sharing of simple models for courselet structures, the modeling of domain knowledge, tasks and user competence as well as personal user preferences in such a way that it can form a basis for course structuring and personalization, the personalization of content modules through annotations expressed as meta-data, the personalization of courses or courselets through different ways to configure and modify structures of modules and the adaption and combination of user interface metaphors for authoring and use of courselets.

In our approach the students will be stimulated to play with preauthored vizualisations and other multimedia explanation modules for mathematical concepts, create, reuse or modify such modules, create their own conceptual models of mathematical knowledge, annotate nodes in conceptual structures with personal information, create courselets based on sequences of explanations generated from the personal conceptual models, indirectly create courselets generated through knowledge-based techniques basing their inferences on meta-data coding relevant contextual information, browse courselet structures, exchange courselets. Conzilla will provide the basic platform for the these activities, and the Graphing Calculator (http://www.pacifict.com) will be the basic visualization tool.

Teachers will also be engaged in the same kind of activities with the purpose to create relevant courselets. Typically the learning situation will be partly teacher driven and partly student driven. The above activities need a set of tools both for authoring (for conceptual modeling and meta- data specification) and browsing. This set will include Conzilla and the Graphing Calculator from the start, but will be incrementally revised during the course of the project. An important part of the project will be the user-testing of Conzilla in a realistic learning environment, and the features of the program will be modified according to the feedback of the participating teachers and students.

From a methodological and technical point of view, the research will combine methods and techniques from the areas of conceptual modeling, ie. design of ontologies, concepts and relationsships suitable for modeling mathematical knowledge, artificial intelligence, i.e. knowledge representation for personal information, human machine interaction, i.e. adaptive interfaces, cognitive science ie cognitive aspects of concept learning.

Deliverables, Timelines, Testbeds The submodule will be focussed on the use of personalized courseware for a few courses in Mathematics on the Information Technology Program at KTH.

In the first half year, we will introduce the Conzilla/GraphingCalculator-based methodology in two mathematics courses, set up an archive of appropriate multimedia explanation modules. Teachers will add appropriate meta-data to explanation modules, following up on studies of students modeling their own mathematical knowledge. Teachers will create prototypical courselets manually using conceptual modeling and semiautomatically using knowledge-based techniques.

In the second half year, students will also use these pre-authored courselets, author their own explanatory module, and be able to browse archives of multimedia explanations. Students will generate personalized courselets based on conceptual models, will share courselets among each other on the same course and will add annotations to modules and experiment with knowledge-based generation of courselets.

In the third half year, we will perform a systematic study of courses where the above activities run in parallell. In the fourth half year, evaluation will be performed, guidelines for the methodology developed will be written down.

Results from the submodule will be prototypes, empirical studies, courselet and content archives for particular subdomains of mathematics as well as general reports.

Collaboration/Scholarly Exchange. Strong collaboration between DSV/KTH and KBS/Hannover with respect to interfaces for personalized course material and knowledge-based techniques to generate coursematerial from coursemodules as described in two of the other submodules.

Connections to the work with meta-data in module on Infrastructure and intelligent services. The empirical results of the meta-data activities of this module will serve as input to the development of new metadata-handling capabilities in the infrastructure module.

Strong collaboration with the Mathemagic project at the Advanced Media Technology (AMT) laboratory at KTH, where a multi-media based component archive with mathematical content is being developed under the coordination of Naeve.

Exchange of graduate students planned between KTH and Hannover.

Budget Overview (incl. overhead costs):

KTH/DSV: 30K first year, 30K second year. Budget will pay for one part-time Ph.D. student, infrastructure costs, travel and exchange.

KTH/CID: 10K first year, 10K second year. Budget will pay for one part-time Ph.D. student, infrastructure costs, travel and exchange.

5.4 Personalized Interfaces and Access to Educational Media

Working Title. Personalized Interfaces, Navigation and Guidance - PerINaG

Contributing Research Groups and PIs. Hannover KBS (Nejdl/Allert)

Problem Description. Using information spaces learners must integrate new chunks of information into a coherent mental representation. This coherence formation process makes great demand on learners‘ cognitive and metacognitive skills. They must make many decisions and there is a huge number of possible routes which can be constructed and performed by the learner. They must orient themselves and build up connections between single concepts, learning objects, units and courselets. They have to relate important items of content. In hypertext navigation and navigation in non-linear data bases learners suffer from conflicting and competing goal intentions as well as from cognitive overload if the navigational task consumes too much of their resources. There is a strong need of distraction and violation protection in learning and problem solving. The instructional design is therefore based on theories of working memory.

By using information spaces in project oriented learning, learners have to perform real tasks such as answering questions, solving problems, writing reports, etc. So they do not only have to find information but also scan, read, interpret, evaluate for utility, annotate, and form coherence. How can learners be assisted most effectively in orientation, access and navigation? What kinds of direct and indirect help/aid should be presented? How to aid navigation in information spaces?

Does it help learners to be presented the underlying complex domain/knowledge model which is organized and structured net-like? Or do they need different access structure? For example: task-adaptive access/navigation structure facing the needs in different stages of a project (broad overview, goal setting, information seeking, performing the procedural tasks, reflection and analyses), depending on their prior knowledge. How to contextualize learning objects and units and to organize access? Can different access-structure face different task related as well as personal needs (e.g. different learning styles as well as navigation styles/more or less goal orientation)?

State-of-the-art work deals with hypertext navigation and conflicting goal intensions (distraction and volitional protection in learning and problem solving [16]) and with coherence formation in learning with hypertext [34] (learners benefit from learning with hypertext corresponding to their goal orientation and prior knowledge). On the other hand research deals with visualizing techniques in a more technical perspective. These visualizing techniques do not reduce complexity for the learner.

Another research field investigates the effectiveness of adaptive interfaces for instructional systems and adaptive link annotation [14], [36] and [43]. This work has produced no positive outcome for adaptive navigation support for a general population of novice learners. Experimental design has to be developed further in this research field, the impact of different factors has to be evaluated. There is evidence for the inference that learners must agree with adaptive navigation help.

Research Plan and Deliverables. This submodule will focus on course structuring, access, navigation, and orientation in non-linear information spaces, as well as instructional and cognitive impacts on the design of distributed learning repositories, as well as requirements for designing these systems with evaluation in mind from the beginning (protocol and trace functionality, educational setup, etc.) It will include applied research based on existing experimental studies in cognitive science, and design of interfaces for knowledge navigation.

In a first step we will investigate the effect of different access-structures. Learners will be able to choose from different linear or hierarchically structured trails: Each learning object (fragment) is located in those trails. Learners are given advice of pre- and post-concepts to be learned. The trails are designed along different learning theories. Metadata is used to generate those learning sequences.

Field Research: We observe the requirements of learners in different stages of project oriented as well as self-organized learning. Testbeds in Germany are described in the Edutella module. Questions: Does navigation in the non hierarchical (meshed, net-like) domain/knowledge model support the process of coherence formation? Or should access structures be different (hierarchies, linear trails, associative relations)?

Task-adaptive access: We will investigate how to contextualize and how to organize access (coping with the different needs of learners in different stages of problem solving, project/problem based learning related to specific tasks within a project-life-cycle). Different trails/paths based on different learning theories will be dynamically generated and the use of these trails will be evaluated. In controlled experiments we will evaluate the effectiveness of different access structures as well as of direct and indirect navigation aids. Results will guide the refinement of tools and design of prototype implementations.

In a second step we will combine these studies with our results within the PLease Module (personalized course sequencing), again including experimental studies in the second year. We will investigate the effect of personalized access and adaptive interfaces (e.g. adaptive link annotation).

Deliverables will also include guidance for teachers, including a catalogue based on existing learning theories, to support teachers and course builders in decision making. These guidelines will reflect various conditions and factors of learners (prior knowledge, domain etc.) and guide them during the design of their courses in special domains. This assistance will not provide presentation recipes, but aids for decision making: how to structure access to learning repositories, which factors to consider in designing hypertext navigation, etc.

These studies will also be embedded in an overall evaluation, based on theory based evaluation as decribed by the WGLN evaluation team [26, 37]. We will (in strong collaboration with PIs) detect and determine appropriate set of goals to focus evaluation, and will use it to guide requirement specifications and refinements. For these, a report will be one of the main deliverables. Field studies and experimental studies will be part of ongoing evaluation.

We will also specify functional requirements within different modules to facilitate evaluation, including logfiles, protocol and tracking functionalities to observe learners´ interaction and movement in the information space/open learning repository. Complexity of action and navigation is one focus of observation.

Dissemination & Testbeds Testbeds: mainly ULI and CS courses at KBS (see Edutella), other testbeds in interaction with collaborating modules. Dissemination: Publications and Reports.

Collaboration and Scholarly Exchange Especially with Edutella, PleaSe, Modular Content Archives, in Sweden with Department of Teacher Education. Cooperation and research visits.

Budget Overview (incl. overhead costs):

KBS: 30K first year, 30K second year, for part-time Ph.D. student, incl. overhead costs, travel and exchange.

L3S central: 15K first year, 15K second year, for part-time Ph.D. student in media and design education, especially for evaluation, incl. overhead costs, travel and exchange.

References

[1] Margareta Westergren Axelsson. USE – The Uppsala Student English Corpus: an instrument for needs analysis. *ICAME Journal*, 24:155–157, 2000.

[2] Lars Borin. A corpus of written Finnish Romani texts. In *LREC 2000. Second International Conference on Language Resources and Evaluation. Workshop Proceedings. Developing Language Resources for Minority Languages: Reusability and Strategic Priorities*, pages 75–82. ELRA, 2000.

[3] Lars Borin, editor. *Parallel corpora, parallel worlds. Papers presented at a symposium on parallel and comparable corpora, Uppsala, 22–23 April, 1999*. Rodopi, Amsterdam, t a.

[4] Lars Borin and Mats Dahllöf. A corpus-based grammar tutor for education in language and speech technology. In *EACL'99, Computer and Internet Supported Education in Language and Speech Technology. Proceedings of a Workshop Sponsored by ELSNET and The Association for Computational Linguistics*, pages 36–43. ACL, 1999.

[5] Lars Borin, Kariné Åkerman Sarkisian, and Camilla Bengtsson. A stitch in time: Enhancing university language education with web-based diagnostic testing. In *Proceedings of the 20th ICDE Conference, Düsseldorf 1–5 April 2001*. ICDE, t a .

[6] Lars Borin, Kariné Åkerman Sarkisian, Camilla Bengtsson, and Monica Langerth Zetterman. Developing and evaluating web-based diagnostic language testing in university language education. In *Accepted for the ALTE European Year of Languages conference, Barcelona, 5–7 July, 2001*. ALTE, t a .

[7] Don Box, David Ehnebuske, Gopal Kakivaya, Andrew Layman, Noah Mendelsohn, Henrik Frystyk Nielsen, Satish Thatte, and Dave Winer. Simple object access protocol (soap) 1.1. Technical report, W3C Note, May 2000. http://www.w3.org/TR/2000/NOTE-SOAP-20000508/.

[8] Jeen Broekstra, Michel Klein, Stefan Decker, Dieter Fensel, and Ian Horrocks. Adding formal semantics to the web: building on top of RDF schema. In *Proceedings of the 10th World Wide Web Conference*, Hongkong, May 2001.

[9] P. Brusilovsky. Methods and techniques of adaptive hypermedia. *User Modeling and User Adapted Interaction*, 6(2-3):87–129, 1996.

[10] Peter Brusilovsky. Adaptive and intelligent technologies for web-based education. *Special Issue on Intelligent Systems and Teleteaching, (Künstliche Intelligenz),*, 4:19–25, 1999.

[11] Stefan Decker, Michael Erdmann, Dieter Fensel, and Rudi Studer. Ontobroker: Ontology based access to distributed and semi-structured information. In R. Meersman, Z. Tari, and S. Stevens, editors, *Semantic Issues in Multimedia Systems*. Kluwer Academic Publisher, Boston, 1999.

[12] Stefan Decker, Frank van Harmelen, Jeen Broekstra, Michael Erdmann, Dieter Fensel, Ian Horrocks, Michel Klein, and Sergey Melnik. The semantic web - on the roles of XML and RDF. *IEEE Internet Computing*, September 2000.

[13] Rael Dornfest and Dan Brickley. The power of metadata. In Andy Oram, editor, *Peer-to-Peer: Harnessing the Power of Disruptive Technologies*. O'Reilly, 2001. http://www.openp2p.com/pub/a/p2p/2001/01/18/metadata.html.

[14] John Eklund and Ken Sinclair. An empirical appraisal of the effectiveness of adaptive interfaces for instructional systems. *Educational Technology & Society*, 3(4), 2000.

[15] H. Garcia-Molina, Y. Papakonstantinou, D. Quass, A. Rajaraman, Y.Sagiv, J. Ullman, V. Vassalos, and J. Widom. The tsimmis approach to mediation: Data models and languages. *Intelligent Information Systems (JIIS), Kluwer*, 8(2):117–132, 1997.

[16] Peter Gerjets, Katarina Scheiter, and Elke Heise. Navigation and conflicting goal intentions: Distraction and volitional protection in learning and problem solving. In I. Wachsmuth and B. Jung, editors, *Proceedings der 4. Fachtagung der Gesellschaft für Kognitionswissenschaft*, pages 73–78. Infix, 1999.

[17] Jeff Heflin and James Hendler. Semantic interoperability on the web. In *Proceedings Extreme Markup Languages*, Montreal, August 2000.

[18] Nicola Henze. *Adaptive Hyperbooks: Adaptation for Project-Based Learning Resources*. PhD thesis, University of Hannover, 2000.

[19] Nicola Henze and Wolfgang Nejdl. Adaptivity in the KBS Hyperbook System. In *2nd Workshop on Adaptive Systems and User Modeling on the WWW*, Toronto, Canada, May 1999.

[20] Nicola Henze and Wolfgang Nejdl. Extendible adaptive hypermedia courseware: Integrating different courses and web material. In *Proccedings of the International Conference on Adaptive Hypermedia and Adaptive Web-Based Systems (AH 2000)*, Trento, Italy, 2000.

[21] Nicola Henze, Wolfgang Nejdl, and Martin Wolpers. Modeling Constructivist Teaching Functionality and Structure in the KBS Hyperbook System. In *Proc. Computer Supported Collaborative Learning Conference*, Stanford, December 1999. a previous version appeard in AIED99 Workshop on Ontologies for Intelligent Educational Systems.

[22] Ora Lassila. Web metadata: A matter of semantics. *IEEE Internet Computing*, 2(4):30–37, 1998.

[23] L.Liu and C.Pu. An adaptive object-oriented approach to integration and access of heterogeneous information sources. *Distributed and Parallel Databases*, 5(2):167–205, 1997.

[24] A. Naeve. The garden of knowledge as a knowledge manifold - a conceptual framework for computer supported personalized education. Technical Report CID-17, KTH Stockholm, 1997.

[25] A. Naeve. Conceptual navigation and multiple scale narration in a knowledge manifold. Technical Report CID-52, KTH Stockholm, 1999.

[26] John Nash, Leo Plugge, and Anneke Eurelings. Managing CSCL projects. internal report.

[27] Wolfgang Nejdl and Martin Wolpers. KBS Hyperbook – a data-driven information system on the web. In *8th International World Wide Web Conference*, Toronto, Canada, May 1999.

[28] Wolfgang Nejdl, Martin Wolpes, and Christian Capelle. The RDF schema specification revisited. In *Workshop Modellierung 2000*, Koblenz, April 2000.

[29] M. Nilsson. The conzilla design - the definitive reference. http://conzilla.sourceforge.net/doc/conzilla-design/conzilla-design.html, August 2000.

[30] M. Nilsson and M. Palmer. Conzilla - towards a concept browser. Technical Report CID-53, KTH Stockholm, 1999.

[31] D. Pettersson. Aspect filtering as a tool to support conceptual exploration and presentation. Technical Report TRITA-NA-E0079, KTH Stockholm, December 2000.

[32] T. Risch and V. Josifovski. Distributed data integration by object-oriented mediator servers. In *Concurrency - Practice and Experience J.* John Wiley & Sons, 2001. http://www.dis.uu.se/ udbl/publ/concur00.pdf, to be published.

[33] F. Ruth, J. Tholander, and K. Karlgren. Trusting the tutor - design aspects and trust issues in a prototypical pedagogical assistant. In *7th International Conference on Computers in Education*, Chiba, Japan, November 1999.

[34] Wolfgang Schnotz and Thomas Zink. Information search and coherence formation in knowledge acquisition from hypertext. *German Journal of Educational Psychology*, 11(2):95–108, 1997.

[35] Anne Morgan Spalter and Rosemary Michelle Simpson. Reusable hypertext structures for distance and JIT learning. In *Hypertext 2000*, pages 29–38, 2000.

[36] Markus Specht. Empirical evaluation of adaptive annotation in hypermedia. In *Proceedings of ED-MEDIA & ED-TELECOM98*, volume 2, pages 1327–1332, Charlottesville, VA, 1998. AACE.

[37] Helge Strömdahl and Monica Langerth-Zetterman. On theory-anchored evaluation research of educational settings especially those supported by information and communication technologies. internal report.

[38] J. Tholander, K. Karlgren, F. Rutz, and R. Ramberg. Design and evaluation of an apprenticeship setting for learning object-oriented modeling. In *7th International Conference on Computers in Education*, Chiba, Japan, November 1999.

[39] J. Tholander, F. Rutz, P. Johannesson, K. Karlgren, and R. Ramberg. A pedagogical assistant for learning object-oriented design: Nagging students into self-reflection. In *Proceedings of the 10th International PEG Conference*, Exeter, England, July 1999.

[40] A. Tomasic, L. Raschid, and P. Valduriez. Scaling access to heterogeneous data sources with DISCO. *IEEE Transactions on Knowledge and Data Engineering*, 10(5):808–823, 1998.

[41] V.Josifovski and T.Risch. Functional query optimization over object-oriented views for data integration. *Intelligent Information Systems (JIIS)*, 12(2–3):165–190, 1999.

[42] V.Josifovski and T.Risch. Integrating heterogeneous overlapping databases through object-oriented transformations. In *Proc. 25th Conf. on Very Large Databases (VLDB'99)*, pages 435–446, Edinburg, September 1999.

[43] G. Weber and M. Specht. User Modeling and Adaptive Navigation Support in WWW-Based Tutoring Systems. In *Proceedings of the Sixth International Conference on User Modeling, UM97*, Sardinia, Italy, 1997.