

# Community Building around Encyclopaedic Knowledge

---

Josef Kolbitsch<sup>1</sup> and Hermann Maurer<sup>2</sup>

<sup>1</sup> Graz University of Technology, Austria

<sup>2</sup> Institute for Information Systems and Computer Media, Graz University of Technology, Austria

This paper gives a brief overview of current technologies in systems handling encyclopaedic knowledge. Since most of the electronic encyclopaedias currently available are rather static and inflexible, greatly enhanced functionality is introduced that enables users to work more effectively and collaboratively. Users have the ability, for instance, to add annotations to every kind of object and can have private and shared workspaces. The techniques described employ user profiles in order to adapt to different users and involve statistical analysis to improve search results. Moreover, a tracking and navigation mechanism based on trails is presented.

The second part of the paper details community building around encyclopaedic knowledge with the aim to involve “plain” users and experts in environments with largely editorial content. The foundations for building a user community are specified along with significant facets such as retaining the high quality of content, rating mechanisms and social aspects. A system that implements large portions of the community-related concepts in a heterogeneous environment of several largely independent data sources is proposed. Apart from online and DVD-based encyclopaedias, potential application areas are e-Learning, corporate documentation and knowledge management systems.

*Keywords:* Digital Libraries, Electronic Encyclopaedias, Knowledge Brokering Systems, Active Documents, Annotations, Knowledge Management, Tracking, Adaptation, Community Building.

## 1. Introduction

Numerous electronic encyclopaedic systems are available, however most are rather conservative, inflexible and static. Their focus is on technical aspects such as data storage, retrieval and links to the Internet.

The following sub-section gives a brief overview of the capabilities of traditional electronic encyclopaedias; sections 1.2 and 1.3 introduce two popular state-of-the-art encyclopaedias, and sections 1.3 and 1.4 address substantially different approaches to encyclopaedic knowledge.

### 1.1. Traditional Electronic Encyclopaedias

Electronic Encyclopaedias were introduced in the 1980s (e.g., [15]). Early versions were published on CD-ROMs in order to host the large amount of data. In the 1990s the Internet was discovered as a way to offer access to electronic encyclopaedias. The approach of most electronic encyclopaedias available on the market is rather conservative in that they focus on simplicity of information retrieval, embedding into applications such as word processors, links to search engines on the Internet, inclusion of multimedia content (images, videos, 3D animations), etc.

As such, the typical electronic encyclopaedia is a digital collection of material that used to be available in a series of books. Currently, the benefits of a digital representation are mainly ease of search and retrieval, that they are relatively easily kept up-to-date, that they are available anytime anywhere, and that they can offer media such as animations or video clips. However, many things people usually do in printed dictionaries and encyclopaedias cannot be done in their digital counterparts. Users cannot make annotations, do not have the possibility to highlight parts of the content or to make clippings, etc.

More importantly, digital encyclopaedias can open up entirely new perspectives and modes of application. The power of distributed environments, for instance, and the potentiality for users to collaborate are still mostly ignored.

## 1.2. Brockhaus Multimedial

The “Brockhaus Multimedial 2005 Premium” Encyclopaedia [5] is undoubtedly one of the most advanced electronic encyclopaedias currently available. Apart from the functionality of conventional electronic encyclopaedias, the Brockhaus Multimedial includes several innovative features such as the “knowledge network” (see [29; 30]). The software finds similarities and connections and graphically depicts them in an intuitive way. The encyclopaedia also offers a large number of links to external articles, to a large image database, to specialised resources on the Internet, and to popular general-purpose search engines.

## 1.3. Wikipedia

The Wikipedia is the prime example for a free, Web-based, community-driven encyclopaedia based on Ward Cunningham’s idea of wikis (see [41] and [25]). The content of the Wikipedia can be edited anytime by anyone on the Internet, i.e., every reader can also be an author.

The project is growing rapidly: as of July 2006, Wikipedia is available in more than 200 languages with about 4.6 million articles in all languages combined. The largest editions are the English one with more than 1.2 million and the German one with almost 420,000 articles (source: [42]). The success of the project builds on the concept of tight involvement of all users, the sense of community, and the intent to watch each other and look after the articles.

The methodology and architecture of the Wikipedia make it much more flexible than a print version or an edited online version of an encyclopaedia. Content can be added faster and when it is needed, e.g., when the space mission to Titan was in the news, the Wikipedia article on Titan was largely extended and supplemented with links to the web-sites of Nasa and ESA and the most current photos available. Moreover, Wikipedia can contain articles that are not suitable for edited encyclopaedias such as extensive information on TV series, video games, black magic, sexual practices or conspiracy theories.

The open structure of the Wikipedia is prone to a number of complications, though. Vandalism—deleting articles, appending incorrect or inappropriate information, inserting vulgarities, etc.—and repair, for instance, is only one phenomenon to be observed in Wikipedia (e.g., [39]). Another example are edit wars, where certain paragraphs of an article are repeatedly inserted and deleted by two users or groups. Also antisocial behaviour and “identity fraud” of a small group of users are pestering the Wikipedia community.

One of the aims of Wikipedia is not to be biased; this concept is named “neutral point of view”. However, even if an article is written in a neutral way the different

cultural, social, national and lingual backgrounds can have an enormous influence. As of February 7th, 2005, the English article on the American chess player Paul Morphy, for instance, has 5466 words and contains a photo, citations and references to external resources. In contrast to this, the article in the German version merely consists of 290 words and does not offer any additional information. Also, the English edition of the Wikipedia contains relatively long articles on various TV series, whereas the pages on topics such as Asian art or political events in Russia are very concise. So even if articles are written objectively, there is a particular imbalance and bias due to the environment and experience of the authors.

In addition to this, mechanisms to approve the expertise of authors or to verify the authenticity of descriptions do not exist. Even more importantly, research has shown that content does not stabilise, i.e., articles are modified again and again, do not converge in size and tend to grow steadily (see [39]). Thus, articles in the Wikipedia are not useful as reference or for quotations because text might be changed significantly or even removed from articles.

Although it is beyond the scope of this paper, it should be mentioned that the Wikipedia project has a potentially deep “philosophical” impact. In the modern age, the tradition to trade knowledge in the form of professionally authored encyclopaedias goes back to the 17th and 18th centuries. This is in stark contrast to the Wikipedia, where articles are neither written by experts nor are they reviewed by editors, there is no fact-checking, and the content is freely available. The principle of the Wikipedia stems from the open source community: work in a community is done to meet one’s own needs, for the good of the public and for free. (See also GNU Free Documentation License, [14], and Creative Commons Project, [10].)

## 1.4. Blogs

Weblogs, often called blogs, are web-pages that contain newsgroup-like articles in a reversed chronological order. They are updated frequently, typically once a day, and can be maintained by a single person, a group of individuals or even by the public (e.g., [4]).

Although most blogs are employed for personal use (as a sort of diary, where people can add their comments, for example), businesses begin to discover their value in a corporate background (see [38]). In most cases, they are used strictly inside companies in order to keep co-workers up-to-date on the progress of a project, to discuss certain topics in an informal way, etc. Sometimes, however, they are also used externally to ignite discussions between developers of a product and the consumers.

In their nature, blogs resemble wikis as they offer users a way to express their opinions on certain topics, share

their experiences and knowledge. The main difference is, though, that users are not able to modify articles once they are posted. Thus, information in a blog is fairly persistent.

### 1.5. Google Answers

A radically different approach to encyclopaedic knowledge is taken in knowledge brokering environments such as Google Answers ([16]). In this system, users choose a problem domain, ask a specific question, and determine how much they are willing to spend on an answer. Domain experts reading the question can post an appropriate answer along with references and other resources. They are remunerated for every satisfactory answer. Also registered users have the possibility to submit answers; they will not be paid, though. Questions and their corresponding answers are archived in the system, and other users can browse the repository.

The question “*Why is the sky blue?*”, for instance, is answered by a registered user. Later, a domain expert, a physicist, appends a detailed explanation of the phenomena leading to the blue colour of the sky. The question and both answers are stored in the “geophysics” category of the system and are available to all users.

Although a knowledge brokering system largely deals with encyclopaedic knowledge there are fundamental differences to conventional encyclopaedias: the information is not structured, does not necessarily have to be persistent, its authenticity cannot be guaranteed, etc. However, both experts and registered users attempt to maintain a high quality and professional style.

Thus, knowledge brokering systems like Google Answers are largely community-driven, and quality as well as content depend on experts and users participating in the project.

## 2. Functionality Beyond Traditional Electronic Encyclopaedias

The approaches to electronic encyclopaedias described above are too static, offer too little flexibility, do not support users in their work, and do not open new prospects as they could. In order to counter these deficiencies, seven distinct demands that are essential in an elaborate electronic encyclopaedia system are pointed out in the following sub-sections.

It is quite surprising that despite the fact that these technologies have been employed in hypertext and digital libraries for years, they have not yet been introduced in the field of electronic encyclopaedias. We

believe that together with the collaborative functionality described in section 3, truly new applications can be realised.

### 2.1. Annotations

People working with physical encyclopaedias, and books in general, are used to highlighting sections of text with markers, writing down comments with pencils or attaching post-it notes (see [26]). Knowledge management systems, modern digital libraries and journals such as the Journal of Universal Computer Science offer support for annotations (see [20] and [22]).

A modern electronic encyclopaedia should allow users to “annotate everything”: the textual content of an article, images, video clips, sound files, other users’ contributions (see section 3 below), even the links to external references, etc. Users should also have the possibility to highlight certain portions of text or mark sections of images, videos or sound files. Moreover, users should be able to attach (potentially varicoloured) labels such as “important” or “for project A” to all kinds of objects.

In order to support collaboration among users annotations can be on different levels of access: private annotations are only available to the one user, group annotations can be seen within a specified group of users, and public annotations are visible for everyone. Private and group annotations may simply be created by users themselves, whereas public annotations need some kind of regulations by administrators of the system in order to prevent confusion (see [28]).

### 2.2. Active Multimedia Documents

The idea of active documents is that users can ask arbitrary questions to documents, and answers are provided immediately and seemingly by the document itself (see [18]). The implementation of this concept includes an “online” and an “offline” component. If a semantically equivalent question can be found in the system its answer is presented to the user. Thus, the answer is provided online.

If an appropriate, existing question cannot be retrieved the user’s request is forwarded to a human expert, and the user gets an apologetic message that an answer will be provided as soon as possible. In this case, the answer is provided offline by an expert. In the course of time, typically after some 500 to 1,000 users per document, answers are available for the most significant and most frequently asked questions. Therefore the human experts are no longer required, and answers can be provided by the online component.

In an environment that handles encyclopaedic knowledge, it should be possible to ask questions to

every piece of content, i.e., most objects in the system are active documents. Answers are provided either by an editor or by a user of the system (see section 3 below). A large number of questions to an editorial article causes a notification to be sent to the respective editor, and the article has to be reviewed. If too many questions to an article written by a user from the community occur the article is (at least temporarily) disabled and not visible to other users.

### 2.3. Links to External Resources

An electronic encyclopaedia should be massively hyper-linked—both internally and to external resources. External resources do not only comprise general-purpose search engines such as Google and specialised databases such as the Internet Movie Database, IMDb, or a dedicated encyclopaedia on chemistry but also links to other kinds of (relatively) persistent information such as newspapers, magazines and journals. With these resources, an encyclopaedia can offer features such as “in the news”. Moreover editorial articles from the encyclopaedia can be supplemented with high-quality, up-to-date information from the news. Content on the Ukraine, for example, could be complemented with a report from the BBC on the recent election in the country.

The inclusion of external resources, however, poses interesting problems such as the consistency of links and quality control. Link consistency is relevant internally as well as externally (cf., [21]). A hyperlink to the current programme of the Globe Theatre in an article on London has to be removed, for instance, when it becomes unavailable. On the other hand, it has to be ensured that bookmarks users make to articles in the system can be accessed even if the article has been modified. Persistent URLs might offer a partial solution to this problem (e.g., [33]).

When external resources are offered quality becomes an issue as well. Hence, techniques such as blacklisting and whitelisting might have to be employed (see section 3.3 below).

### 2.4. Private and Shared Workspaces

Electronic encyclopaedias should allow for the users’ demand to retain clippings of articles, other pieces of content and information (e.g., [27]). Therefore private and shared workspaces are proposed for electronic encyclopaedias by which a central archive for storing material is provided. In order to make use of the visual memory and pattern recognition skills of humans a two-dimensional workspace similar to a spatial hypertext seems favourable (see [8])

Personal workspaces may contain textual content as well as pictures, video clips, animations, etc. Users can either copy and paste the material they discover or,

alternatively, they can extract parts thereof by way of transclusions (see [31; 32] and [23]). In the latter approach, the information is not duplicated but the original data is “cited”, i.e., every time a transcluded piece of information is accessed, it is retrieved from the corresponding, original document. Moreover, users can make annotations to objects in their personal workspaces, can save hyperlinks to external resources, can store search queries, can put external documents into their workspace, can start to assemble new information and perform similar tasks. The personal workspace also provides an overview of articles recently read and a history of most recent comments and discussions.

Shared workspaces are essential for collaboration in user groups (e.g., [37]). They provide the basic functionality of personal workspaces but have to take access control and communication facilities into consideration. To the benefit of ease of use, only two access modes are provided: private (to the author) and public (to the group).

Functions to support communication among group members in a shared workspace are simplistic: a user can leave a note for one user or for the entire group. When a user is actively working in the workspace and receives a message it is displayed instantly; otherwise it is presented as soon as the user returns to the workspace (see also section 3.6 below).

### 2.5. User Profiles and Adaptation

Electronic encyclopaedias should adapt to their users in order to provide better search results in shorter time and to facilitate the discovery of knowledge that might otherwise not have been found. Adaptation to the user allows for differing levels of experience and knowledge as well as distinct interests and varying aims. Students doing research for their homework need to be confronted with different tools and need to get different search results than a teacher, for example, who is looking up a particular detail of the colonisation of New Zealand. Therefore adaptation has to take place in various aspects of an encyclopaedic application: from the user interface and the tools provided to the presentation and level of detail of search results (content-level adaptation) and the material supplied (link-level adaptation; e.g., [7]).

User profiles are utilised to create an adaptive environment. Several differing strategies are suggested for generating a profile: submission of a general profile, creation of an ad hoc profile, or dynamic profiling. A general user profile, for instance, can be filled out by the user on the first use of the encyclopaedia. It contains general questions about the user’s experience and interests in certain domains such as geography, history, literature, science, etc. The type of user interface—

ranging from a simplistic, Google-like search window to an expert mode—can be selected as well (e.g., [6]).

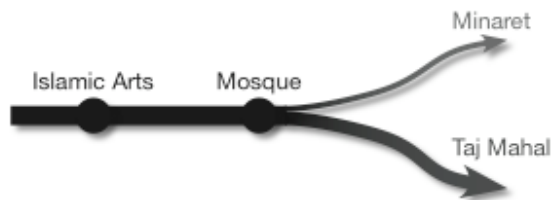


Figure 1: A trail leading from “Islamic Arts” to the term “Mosque”. Suggestions for further steps are “Minaret” and “Taj Mahal” (thicker, i.e., more popular).

Both as an alternative and in addition to the general profile, users can employ ad hoc profiles. The encyclopaedia could ask the user a question such as “*What are you searching for today?*” Users select a problem domain, subsequently determine what types of material they are interested in, and define a search query. The system combines the ad hoc profile with the general user profile, executes the query and is capable of providing more accurate and more specific articles while filtering content that does not match the user’s profile.

An advanced version could include dynamic profiling and adaptation, which basically means that the user’s actions in the environment, the kind of content displayed, the areas from which articles are retrieved, etc. are tracked. Based on these data, the system can adjust the user interface and display more precise results. The system might also put up personalised messages such as “*Did you know that ...*” or “*You might also be interested in ...*” where appropriate. The encyclopaedia’s featured article of the day can be adapted or recent news items might be pointed out to the user.

## 2.6. Trails

Electronic encyclopaedias should make use of an advanced tracking and navigation technique. The mechanism consists of two parts: a tracking component and a navigation component. The tracking component uses a database to retain weights for every possible connection CAB between any two articles A and B. Whenever a user navigates from A to B the weight of the CAB is incremented by one, i.e., paths with higher weights are used more frequently, which indicates higher relevance.

The connection weights are utilised by the navigation component in order to suggest articles to the user (cf. [9; 12; 13]). The trail shows the articles the user has displayed so far (see figure 1). From the second article onwards, the trail also includes the two most relevant (most popular) articles from the perspective of the

current article. The user can navigate to these potentially interesting pages by simply clicking on the trail.

The maximum number of connection weights to be stored is  $n*(n - 1)$ , with  $n$  being the total number of articles. In order to improve the appropriateness of the articles suggested in the trail, the length of the tracking paths can be extended from two to three or even more nodes. However, longer paths consume more space in the database.

It should be noted that adaptation, in fact, also can take place in conjunction with trails. The weights of the tracking paths can be multiplied with information from the user’s profile in order to influence the ranking of the suggested articles. In the example depicted in figure 1, for instance, a user interested in arts might have different topics recommended than someone specialising in the history of Islam.

## 2.7. Quality Feedback and Review Mechanism

Electronic encyclopaedias should retain detailed statistics on the users’ queries and the articles retrieved in order to enhance the performance of the system. The findings derived from the system logs can be employed to find out which articles are currently very popular and, consequently, might have to be revised or extended. On the other hand, the statistics might reveal that certain topics are requested rarely and, therefore, do not need to be updated any more.

In addition to this, the statistical analysis can point out missing articles and instantly notify an editor. Alternatively, the system can offer a function for users to request a new article explicitly. This relatively simple technique can improve the flexibility and topicality of edited encyclopaedias (cf. section 1.3 above).

## 3. Community Building in Electronic Encyclopaedias

As pointed out above, traditional applications handling encyclopaedic knowledge are rarely based on user communities. Three exceptions, though, manage to be quite successful: Wikipedia, specialised blogs and knowledge brokering systems such as Google Answers. However, these systems have drawbacks such as the lack of authenticity of the content provided or the unknown identity of the authors of information.

Therefore a novel concept is introduced that allows a community to build around editorial encyclopaedic knowledge. The following sub-sections address the purpose and notion of the concept, detail foundations of a user community as well as social issues and concerns

about the quality of the content. Furthermore discussion and communication facilities along with techniques to present the content are briefly outlined.

### 3.1. Purpose and Notion of a User Community

In an encyclopaedic environment, users should not only have the capability to read articles and comment on them, but they should be able to contribute actively to the development of the “knowledge base”, to keep it up-to-date, and to make the environment more flexible. Thus, the community built around an electronic encyclopaedia is an opportunity for users to share knowledge and experiences in various ways.

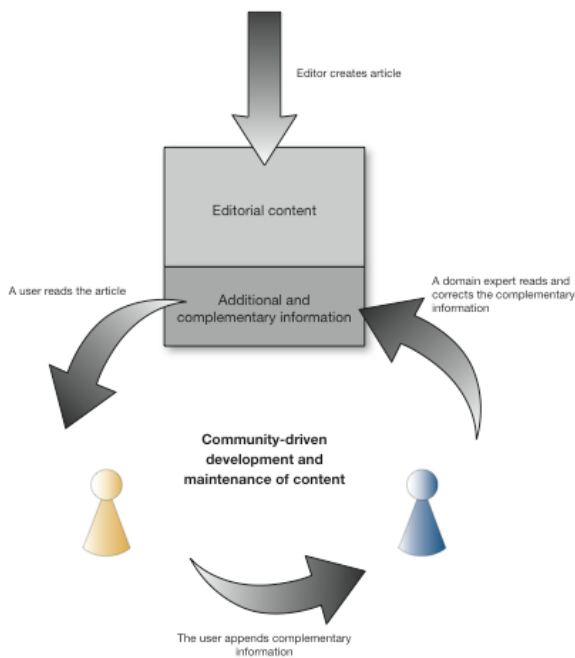


Figure 2: Community-driven development and maintenance of content. Editors create an article and users can add complementary content, references, etc.

Users may not only read but also append information to existing pages or write entirely new articles, can include or reference external resources, and have discussions. A community enables collaboration, so users can actually work *together* on a given project.

### 3.2. Foundations of a User Community

As the focus of the user community is to work on the knowledge base, an emphasis has to be put on the quality of the articles. In order to avoid problems known from the Wikipedia—users cannot be sure whether the article they read is correct—at least the editorial content has to be authentic.

Hence, editorial content is provided by the board of editors and cannot be modified by the user community. When users from the community read an article and would like to supply additional content such as text, photos or sound they may do so. Subsequent users can provide yet more content or change the information that was provided by other users. This “cycle” of community-driven content development is depicted in figure 2.

The concept is based on three levels of users: certified experts, domain experts and plain users (see figure 3). Certified experts are typically experts in their domain (e.g., university professors or researchers) that are certified either by an editor or by a certain number of other certified experts. Potentially new experts can submit an application or can be recommended by other users. Usually their application is discussed by a group of experts in the same domain, and they have to pass an “entrance exam”, in which their expertise is to be proved.

Certified experts have the ability to make a plain user a domain expert when the user, for example, posts a number of high-quality contributions. Plain users can also become domain experts when they have a very large number of positive ratings (see also section 3.2). On the other hand, certified experts or domain experts unable to meet the required standards will have their certification withdrawn.

In an attempt to highlight the community’s professionalism, domain experts and especially certified experts are encouraged to provide a publicly available web-page with personal information, documentation of their expertise (e.g., a list of publications) and similar data.

The combination of a certification mechanism and providing personal and professional information is capable of ensuring high quality standards and can keep vandalism as observed in the Wikipedia low (see [39; 42]) while it still remains easy to submit information to the system. Another measure to foster development of content and deter malice are the quality control functions introduced in section 3.3.

### 3.3. Quality Control

The quality control mechanism is based on blacklisting, certification by experts, and a rating system similar to those known from discussion forums or online auctioning systems. Blacklisting means that the system retains a list of certain words or phrases, for example vulgar expressions, that must not be contained within articles published by users of the community (see [22; 40]). If an article includes a blacklisted expression, it cannot be submitted to the system. Since blacklisting can basically only be applied to all kinds of textual content, it is used for text-based content, annotations, comments, hyperlinks, in discussion forums, etc.

This approach has obvious drawbacks: blacklisting the word “sex”, for example, almost automatically excludes articles about Freud. Therefore for various kinds of information, such as links to external resources, whitelisting is probably better suited (see [24]). This means that only certain, well-chosen external resources can be accessed, and all other sources are rejected.

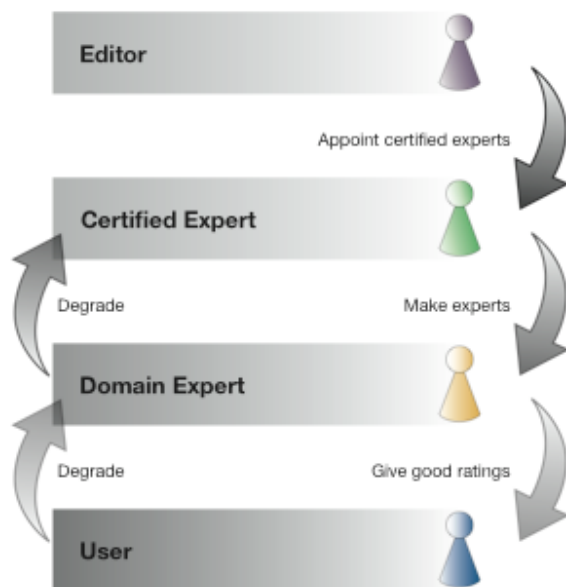


Figure 3: Different levels of users and basic mechanisms to grant and retract user statuses. Certified experts can, e.g., make users experts, and by means of poor ratings users can demote experts.

The second advance to quality control includes the intervention of experts. Every article has to fall into a particular category. Whenever a contribution is published by a plain user or by a domain expert, it has to be approved either by a certified expert or by an editor working in the corresponding category.

Although this technique is certainly rather expensive, it fuels the dialogue between experts and authors. An expert could, for example, suggest to the user that a certain section of the article be changed so that it can be appended to the content. Thus, consensus needs to be reached prior to the publication of a new article, which is, in a way, similar to the review process in scientific publishing.

The third measure is an advanced, credit-based rating system. Credits can be positive, negative or neutral and can be given for any kind of information including text, images, annotations and hyperlinks from both editors and users of the community. Credits from different users have different importance: a positive credit from a certified expert might be worth a value of 1.2, whereas the credit of a domain expert has a value of 1.1, and the credit of a plain user is 1.0.

The overall rating for each user is calculated in an accumulative manner. Five positive credits by certified experts and three positive credits by domain experts, for example, result in an total rating of 9.3 ( $5 * 1.2 + 3 * 1.1$ ). If users receive a very large number of positive credits, the value of their own credits might be increased. The credit of a plain user A, for instance, is increased to 1.05 if A’s total rating exceeds a certain threshold value. Every credit given by A is worth a value of 1.05 from that point onwards. In the case where A’s total rating deteriorates, the value of A’s credits decreases as well.

Although this approach is somewhat complex, it represents a “fairer” rating system that takes the users’ commitment into consideration. Moreover, it encourages users to contribute valuable content because it potentially makes them more “powerful”.

Ratings have an influence on both the users’ articles and their user level. If plain users, for instance, receive a great number of positive ratings they might be promoted to domain experts. If domain experts, on the other hand, receive a lot of negative credits, they might be downgraded to plain user level. Ratings can also trigger certain functions on content (see also section 3.5):

- if a user-authored article receives a large number of positive credits, the encyclopaedia system automatically suggests to an editor to make the article part of the editorial content;
- if the number of negative credits for a user-authored article exceeds a certain threshold it will be automatically removed from the system;
- if certified experts or domain experts have a certain number of articles removed due to negative ratings their certification might be withdrawn;
- if editorial content receives too many negative credits the board of editors is notified and the article has to be reviewed.

### 3.4. Active Knowledge Brokering

Active knowledge brokering is an essential component in a community environment. It lets users choose a particular domain and ask arbitrary questions within this field of knowledge. Answers to the users’ questions are provided either immediately by the system or by experts and editors within a certain period of time.

The concept combines multiple technologies such as natural language queries similar to [2] or [17], syntactic, semantic or heuristic analysis of questions, clustering, active document filters, and conventional knowledge brokering as detailed in section 1.5. An overview of the paradigm is given in figure 4.

User start by selecting a certain area of interest such as “astronomy” and formulate an arbitrary question in

natural language, for example, in plain English. The question triggers a database query in the internal knowledge repositories. The results from the various sources are combined, clustered and displayed to the user as “articles from the archives”.

However, the user’s question also triggers an active documents and answer brokering component. First, an analysis of the question is performed (see [19]). If semantically equivalent questions that have already been answered by human experts are available the user is asked “*Did you mean ...?*” If this active documents filter fails, i.e., none of the suggestions is satisfactory or none are obtainable, the question is passed on to a human expert. In the case where an appropriate answer cannot be provided by domain experts within a certain time, the question is passed on to a certified expert. If an answer still cannot be provided, it is forwarded to an editor.

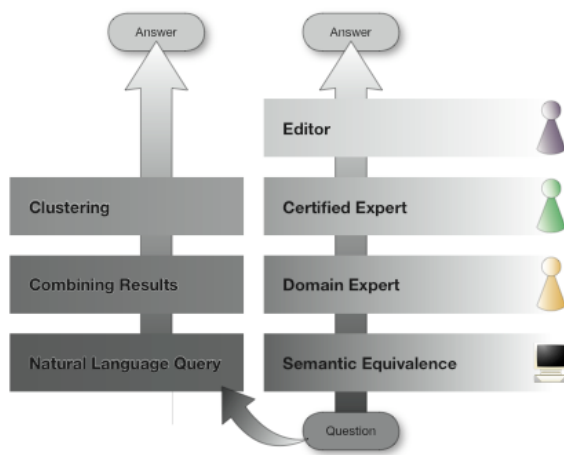


Figure 4: Answer brokering mechanism. When a question is asked the system searches for semantically equivalent previous queries. If none can be found, the question is passed on to domain experts, certified experts and, finally, to editors.

At any point, questions can be rejected as inappropriate. Moreover, experts and editors have the capability of combining semantically equivalent questions that are not recognised by the system. So if a question is similar to a prior one, it is specified as “equivalent” to an existing question and linked to the corresponding answer. Thus, the semantic filter is able to learn.

From a technical perspective, active knowledge brokering is a generalisation of the active documents approach. With active documents, question-answer pairs are always attached to particular articles or other pieces of information. Active knowledge brokering, on the other hand, is question-centered: questions are largely autonomous and are usually not attached to articles. An article or a piece thereof, however, can be part of an answer. Apart from that, an answer to a question can, by means of trans-clusions, lead to a new article.

### 3.5. Social Aspects

Research has shown that at least two types of users exist in environments where digital content is shared and disseminated: “givers” and “receivers” (see [27]). This implies that certain users are willing to update incorrect articles, answer other users’ questions, or even collaborate on certain issues. Other users, in contrast to this, will rather receive information than make an effort to reply to questions, etc.

In order to encourage *all* users to contribute to the community—and in addition to the rating system—these socially disparate characters should be visualised in the environment. A straightforward approach is the use of distinct icons for different types of users and characters. A teacher, for instance, can be depicted as a man or woman in front of a blackboard. A teacher writing many articles has a pile of notes in the icon, and a professor answering a lot of other users’ questions has a warm smile, for instance.

### 3.6. Mixed and Split Content

An article on the history of film, for example, has a large number of annotations and links to external references attached. In order to display such an extensive page in a comprehensible yet intuitive way, two modes for displaying articles are proposed: split and mixed content modes. In split content mode editorial content is strictly separated from user authored content. As depicted in figure 5 (c), the editorial section of an article is always displayed on top. Complementary information added by users is displayed below in an individual area of the window; it can be revealed and hidden by clicking on the triangle symbol. This example also shows two additional sections that are “collapsed”: discussions and external resources. Instead of collapsing areas in the window, icons can be used to indicate additional content such as annotations or hyperlinks. By clicking on an icon the corresponding annotation or hyperlink is revealed.

The main benefit of the split content approach is that it is apparent for users which content was published by editors (and therefore is authentic) and which content was added by other users of the system. Moreover this functionality can be used easily and is straightforward to implement.

A more innovative technique for displaying articles is the mixed content mode. When an expert user, for instance, wants to add a paragraph to an editorial article the entire article is presented in a text editor. The expert can add content including text, images and video clips at any position in the existing article. The editorial content, however, is protected and cannot be changed. Users viewing the article can choose whether the user authored content is displayed (see the checkbox in the title bar of figures 5 (a) and (b)). In figure 5 (a), the

piece of text provided by the community is displayed using a different font.

This approach ensures that the quality of editorial content cannot not be tampered with, while users of the community can easily add material at the most appropriate position in the article, i.e., the context is preserved. This can result in a seamless integration of editorial and community-provided content. Readers, on the other hand, can choose to view only the editorially approved information (figure 5 (a)) or can also have the data supplemented by the community displayed (figure 5 (b)).

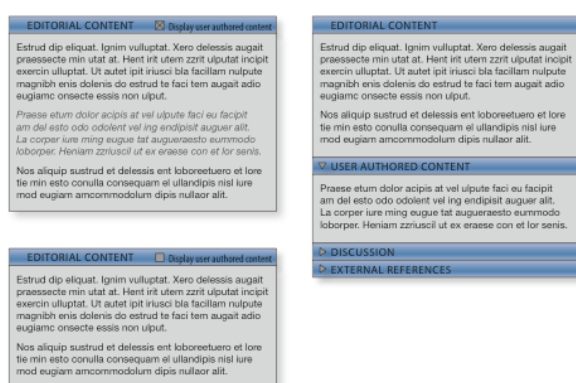


Figure 5: Mixed and split content modes. (a) Left, top: mixed mode with user authored content. (b) Left, bottom: mixed mode without user authored content. (c) Right: split mode with user authored content displayed and discussions and external references hidden.

Together with statistical analysis functionality introduced in section 2.7 above, the concept can even be refined: articles that contain a lot of data provided by the community are automatically reported to the board of editors. When the article is reviewed by editors they can simply accept it, i.e., the information provided by a user is automatically added to the editorial content and credited to the corresponding user. The user is not only associated with the section in the article but also gets an extra positive rating or a particular reward.

### 3.7. Notification and Communication among Users

Notification mechanisms and communication among users are two significant aspects in the community. Notification mechanisms are necessary to let users know that one of the articles they authored was certified or that information has been added to an area of interest. In both cases, either “pull” or “push” technology can be employed. Push technology implies active dissemination where the encyclopaedia sends a message, e.g., an e-mail, to the users informing them of the new information. With the pull technique, the user has to retrieve the corresponding information explicitly

from the encyclopaedia. A technology that is very well suited for this kind of information retrieval is RSS (RDF Site Summary, see [34] and [35; 36]).

It is important that direct user-user and user-group communication can take place within the system and does not require external technologies such as e-mail. Therefore a simple system of „leaving notes“ for users is suggested as the primary way of personal communication: when user A sends a message to a user or user group B, and B is online, the message is presented immediately. Otherwise the note is retained in the system, and as soon as B logs on to the system all messages stored are displayed.

A more open way to communicate are discussion forums briefly outlined in the next sub-section.

### 3.8. Discussion Forums

Discussion forums based on the paradigm of conventional newsgroups should be available in any situation and at any point in the environment. Users have the possibility to discuss matters related to content in forums and discussions can, in fact, be attached to any kind of information in the system: to an article, to annotations, hyperlinks, images, etc. Although discussions are usually in relation with a particular piece of information, users are also able to have “meta-discussions” in which general strategies such as the reorganisation of a encyclopaedic category or the need for new experts can be argued.

Discussions take place in the system but are out-of-band rather than directly in the article (see also figure 5 (c) where discussions are presented in a separate area). The advantage of this approach is that articles strictly contain content and do not include discussions about content. Thus readers can mainly concentrate on the actual content and still have the option to join a discussion, if they are interested.

## 4. System Proposal

We propose a web-based system that builds a community around an encyclopaedic body of knowledge consisting of at least one electronic encyclopaedia, several archives of articles from magazines, and newspapers, and potentially other sources of persistent information of high quality. The target group is a fairly closed user group of people wishing to have simple access to in-depth, high quality information on a wide range of topics.

The following sub-sections provide a brief overview of the architecture and the main concepts to be implemented. A first prototype is to be available in December 2005.

## 4.1. Overview

The proposed system is based on several sources of data from potentially heterogeneous architectures. The knowledge repositories contain entries from encyclopaedias, newspaper articles and similar data. Moreover, the contributions of the community—annotations, complementary content, discussions, etc.—are retained in the databases of the system.

Building on these repositories, the integral components of the system are implemented. User management, different levels of users, the rating system and adaptation are essential to the environment (see sections 4.2 and 4.5). Communication facilities and answer brokering (see section 4.6) are further components that are deeply rooted in the system core.

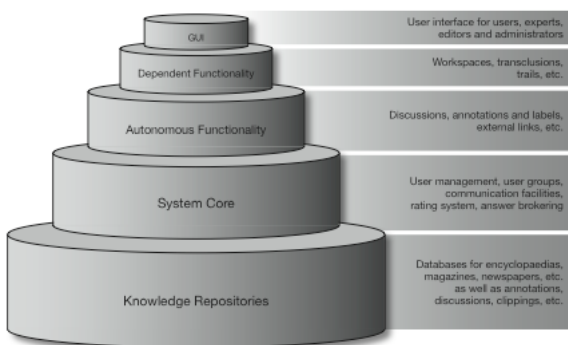


Figure 6: Generic system architecture. The system consists of several knowledge repositories and a system core that encompasses user management, communication, ratings and answer brokering. Building on these foundations, discussion forums and other functionality are realised. Both users, experts, editors and administrators interact with the system in a web-based GUI.

Basic functionality such as discussions, annotations and links to external resources builds on the foundations provided by the system core. More complex concepts such as transclusions or workspaces build on the basic functionality.

The features and functionality offered by the system are presented to editors, experts as well as users in a graphically appealing and easy-to-use, intuitive web-based user interface. An overview of the generic system architecture is given in figure 6.

## 4.2. User Community

The system is based on a user community as introduced in section 3.2 and 3.3. The community can contribute to the knowledge base, develop the content, and work together with the information provided by the system. The system comprises five levels of users:

- administrators that have access to all functions and data in the system, mainly for performing maintenance and organisational tasks;
- editors that may create, modify and delete user accounts as well as all kinds of content in the system, may answer questions, and have full access to discussions; they have also access to the white lists (see below);
- certified experts that can create, modify, remove and certify user authored content, have discussions, answer questions (see below), promote plain users and certify new experts (together with other certified experts); they have access to white lists;
- domain experts that can create new content and modify or remove their own articles, have discussions and answer questions;
- and plain users that can create new content and modify or remove their own articles and have discussions.

All user can rate other users and their contributions.

During a prototype phase, when only a small number of representative users work with the system, the experts in the community are simulated with a set of well-educated people in various areas. Each of these experts is responsible for several domains and has premium encyclopaedias, specialised databases, and similar archives of high quality at hand.

This approach is used to ensure the vividness of the community from the very beginning of the project and to “jump-start” the acquisition of experts and new users.

## 4.3. Basic Functionality

The basic functionality of the proposed system resembles the one of traditional encyclopaedias and digital libraries. It provides full-text search with fault tolerant input over all information stored in the system. In addition to searching, articles can be accessed through browsing.

The system permits users to attach annotations and labels such as “important” or “for my project” to any object in the system. For annotations and labels different access levels are taken into consideration: they can be private, accessible within a certain group of users, or public (see section 2.1).

Furthermore, users are capable of bookmarking articles and any other information items in the system. Thus, they can make direct references to content in the system.

#### 4.4. Authoring New Content, Linking and Transclusions

Users of the community, especially certified experts and domain experts have the ability to generate new content. This is mainly done by authoring new articles, writing annotations, providing links to external resources and having discussions.

New content, however, can also be authored by means of transclusions (e.g., [23]). An expert writing an article on the history of tea might, for instance, want to include a paragraph on the Boston tea party. A brief summary that is well-suited can be found in an article about settlements on the North American west coast of the 18th century. Instead of copying and pasting the paragraph from this article, the user simply transcludes it. Thus, by simply transcluding pieces of information from various articles, completely “new” content can be produced.

Although transclusions generally pose a number of difficulties in web-based environments, their implementation can be made easier because it will only be possible to make transclusions from the internal knowledge repositories of the system. Hence, consistency of transcluded information and potential copyright-related issues can be controlled as well.

An important aspect are links to external resources such as specialised databases. In order to retain a high standard, whitelisting is employed (see section 3.3). The system has a list of external references (servers) that may be used, and all other resources are automatically blocked. Certified experts and editors, however, have the ability to add new, carefully selected resources to the list.

#### 4.5. Adaptation

The system can adapt to its users. The adaptation is based both on a general profile that can be filled out when a user registers with the system and on ad hoc profiles. Dynamic profiling as described in section 2.5 will not be implemented in the prototype of the system but rather in a forthcoming version.

The profiles provided by users are employed for adaptation on both content and link levels. Hence, search results, the links provided, the source of information, the types of user-authored articles displayed, etc. vary based on a user’s preferences. Users have the option to disable the adaptation in order to get a “default view” that would have been used without any adaptation. This ensures that users still have access to *all* information stored in the system and that no content is filtered by the adaptation component inadvertently.

#### 4.6. Clustering and Knowledge Maps

When users search for terms in the encyclopaedia, clustering techniques are employed to combine and display the results in a way similar to [3] and [11]. A query for the term “blue sky” might return a number of results from the encyclopaedia as well as articles from newspapers and other repositories. These results are displayed to the user in clusters such as “atmosphere”, “electromagnetic field”, “pollution” and “space”.

However, clustering is not only utilised in search queries but also when users view articles. Every article displayed is complemented with document clusters that offer a brief overview of pages available on a topic (cf., [1]). This approach facilitates the discovery of information and makes it easier to get an overview of a certain area of interest.

Knowledge maps as mentioned in section 1.2 are basically a graphic depiction of information clusters. Every article in the system is supplemented with a knowledge map on the user’s demand.

#### 4.7. Active Knowledge Brokering

An active knowledge brokering module as described in section 3.4 will be included as a fundamental part of the system. The prototype version will require users to select both a category and a sub-category for their queries, for example, “astronomy” and “solar system”. Subsequently the user can ask a question in natural language.

The system will search the internal repositories, including the encyclopaedia and the newspaper archive, for matching articles. The results are combined, organised in clusters and presented to the user. Additionally, the system will perform a syntactic and heuristic analysis on the question in order to determine if similar questions have been asked before. In case the system is unable to detect appropriate questions-answer pairs from previous queries, it will forward the question to human domain experts of the corresponding problem domain. If the question is not answered within one day by a domain expert it is passed on to a certain group of certified experts. If, after two more days, still no answer is provided the user’s query is forwarded to editors in the respective area. (See also figure 4.)

#### 4.8. Discussion Forums and Communication

Similar to the “annotate everything” paradigm (see section 2.1), every piece of information in the system can be discussed. Discussion forums are typically attached to objects in the system including articles, images and links. However, users can also have meta-discussions that do not have to be related to any specific kind of object (see section 3.7).

The system includes communication facilities as outlined in section 3.6: users have the ability to write messages to other users in the system. The messages are displayed immediately if the recipient is currently online (similar to instant messaging) or presented as soon as the user logs on to the system (similar to traditional e-mail).

In both discussion forums and personal communication, it should be particularly easy for users to include content from articles, make references to objects such as images or annotations, quote contributions to a discussion, etc. Among others, transclusions and link consistency are aspects that will have to be considered.

## 5. Application Areas

The concept of the proposed system will prove to be valuable in numerous fields ranging from e-Learning systems to enhanced electronic encyclopaedias and corporate documentation and knowledge management systems. The following sub-sections briefly outline a few potential application areas.

### 5.1. Enhanced Encyclopaedic Environments

The proposed system clearly targets heterogeneous systems consisting of several data sources such as encyclopaedias, archives of newspapers or magazines. It is obvious that especially communities that have the opportunity to develop and maintain content can be competitive in comparison to the Wikipedia while ensuring a high standard.

However, not only online encyclopaedias but also DVD-based encyclopaedias can make use of parts of the proposed functionality. Annotations, private workspaces, trails as well as user profiles and adaptation can be fully implemented without the need for a connection to the Internet.

### 5.2. E-Learning

Many state-of-the-art e-Learning environments make extensive use of encyclopaedic knowledge, and students utilise electronic encyclopaedias when they are doing their homework or learning for exams. Therefore e-Learning environments are particularly well-suited for the functionality introduced above: students can use private and shared workspaces to work together on projects (see section 2.4), they can make information they encounter available to other students, and are able to use annotations to share their insights. When articles from the encyclopaedia are active documents, students can ask questions to documents and get instant answers.

Discussion forums can be employed to exchange views on articles and the information retrieved from the system, and the notification mechanisms described in section 3.5 keep students informed when a new article or comment are added, or a discussion forum is updated. Lecturers and teachers can, of course, use the same infrastructure as well. They can publish new documents, make annotations to articles, initiate discussions, attach additional content. Thus in this case, teachers (cf. editors in an encyclopaedia) and students (domain experts) form a community.

### 5.3. Corporate Documentation and Knowledge Management

Many companies including computer businesses, the pharmaceutical industry, and car manufacturers offer online support for business customers and consumers. Their support databases basically constitute encyclopaedic knowledge that can be accessed by a closed user group. With the proposed functionality, especially the community building aspects, these corporate documentation and knowledge management networks can be greatly enhanced.

The researcher or developer of the company, for instance, publishes the documentation of a complex software package. Several users attempt to configure the software but run into problems because of a specific computing platform. They can easily make annotations to the original document, start a discussion on the topic, and attach their configuration files.

The business-business area in which several companies collaborate on a common project can make use of the proposed technology as well. In common projects there is a need to share (encyclopaedic) knowledge and work together on it, whereas other pieces of information must not be unveiled. Hence, the community can make knowledge available that is necessary to complete the project without having to make trade secrets public.

## 6. Conclusion

This paper introduced several demands that future systems dealing with encyclopaedic knowledge should meet. Annotations should be omnipresent, and documents should be really active documents. Furthermore, users must have the opportunity to store the information encountered and their comments in a private workspace—and they have to be able to share it with other users in the system. User profiles with the aim of adaptation individual users, statistical analysis and navigation and tracking technologies based on trails will provide more accurate and more detailed search results.

A major topic for future research will have to be community building around editorial encyclopaedic knowledge in order to foster the development of content and tightly involve the user base. Especially new concepts for establishing the basis for such an environment and appropriate processes for maintaining and ensuring the high quality of the material will have to be rendered.

Although the big picture seems to be clear— involvement of users in content development in its broadest sense—numerous aspects have to be investigated in detail. Research has to be done on concepts such as certification of experts or rating systems, and social aspects and phenomena known from the Wikipedia will have to be studied as well. However, also more technical topics need to be addressed: whitelisting, for instance, is not widely used at the moment, and its weaknesses will have to be analysed. Also link consistency, the use and implementation of transclusions, and new approaches such as active knowledge brokering must be discussed in detail.

In-depth studies on the implementation of links to external sources with techniques such as whitelisting and blacklisting, link consistency, and the implications of transclusion in this content will be discussed in an upcoming paper.

## Acknowledgements

The first author would like to thank Edmund Haselwanter for a very constructive discussion on several topics covered in this paper and Christian Gütl for his input.

## References

- [1] A9, <http://a9.com/>.
- [2] AskJeeves, <http://www.askjeeves.com/>.
- [3] Sybase, Inc. A Better Search Engine with Sybase EAServer and Autonomy®, [http://www.sybase.com/content/1034309/EAS\\_CS\\_v2.pdf](http://www.sybase.com/content/1034309/EAS_CS_v2.pdf), (2005) Visited February 20th, 2005.
- [4] Blogger, <http://www.blogger.com/>.
- [5] Brockhaus Multimedial, <http://www.brockhaus-multimedial.de/>.
- [6] P. Brusilovsky Methods and Techniques of Adaptive Hypermedia, *User Modeling and User-Adapted Interaction*, 2-3 (1996), pp. 87-129.
- [7] P. Brusilovsky Adaptive Hypermedia: From Intelligent Tutoring Systems to Web-Based Education. In *Intelligent Tutoring Systems. Lecture Notes in Computer Science* (G. Gauthier et al., eds.), (2000) pp. 1-7. Springer-Verlag Berlin Heidelberg.
- [8] G. Buchanan et al. Integrating Information Seeking and Structuring: Exploring the Role of Spatial Hypertext in a Digital Library, *Proceedings of the 15th ACM Conference on Hypertext and Hypermedia*, (2004) Santa Cruz, CA, U.S.A., pp. 225-234.
- [9] V. Bush As We May Think, *Atlantic Monthly*, 7 (1945), pp. 101-108. See also <http://www.theatlantic.com/unbound/flashbks/computer/bushf.htm>, Visited February 8th, 2005.
- [10] Creative Commons, <http://creativecommons.org/>.
- [11] Clusty, <http://www.clusty.com/>.
- [12] D. DeRoure et al. Memoir – An Open Distributed Framework for Enhanced Navigation of Distributed Information, *Information Processing and Management*, 1 (2001), pp. 53-74. See also <http://eprints.ecs.soton.ac.uk/4240/01/2001-4240.pdf>.
- [13] R. Furuta et al. Hypertext Paths and the World Wide Web: Experiences with Walden’s Paths, *Proceedings of the 8th ACM Conference on Hypertext and Hypermedia*, (1997) Southampton, UK, pp. 167-176.
- [14] Free Software Foundation, Inc. GNU Free Documentation License, <http://www.gnu.org/copyleft/fdl.html>, (2005) Visited February 8th, 2005.
- [15] R. J. Glushko et al. “Hypertext Engineering”: Practical Methods for Creating a Compact Disc Encyclopedia, *Proceedings of the ACM Conference on Document Processing Systems*, (1988) Santa Fe, NM, U.S.A., pp. 11-19.
- [16] Google Answers, <http://answers.google.com/>.
- [17] GuruNet, <http://www.gurunet.com/>.
- [18] E. Heinrich and H. Maurer Active Documents: Concept, Implementation and Applications, *Journal of Universal Computer Science*, 12 (2000), pp. 1197-1202. See also [http://www.jucs.org/jucs\\_6\\_12/active\\_documents\\_concept\\_implementation/](http://www.jucs.org/jucs_6_12/active_documents_concept_implementation/).
- [19] E. Heinrich et al. Learner-Formulated Questions in Technology-Supported Learning Applications, *Proceedings of ED-MEDIA’01*, (2001) Tampere, Finland, pp. 720-725. See also [http://www.iicm.edu/iicm\\_papers/learner\\_formulate\\_d\\_questions.pdf](http://www.iicm.edu/iicm_papers/learner_formulate_d_questions.pdf).
- [20] Journal of Universal Computer Science (J.UCS), <http://www.jucs.org/>.
- [21] F. Kappe Maintaining Link Consistency in Distributed Hyperwebs, *Proceedings of the INET’95*

- Conference*, (1995) Honolulu, HI, U.S.A. See also <http://www.isoc.org/HMP/PAPER/073/html/paper.html>.
- [22] H. Krottmaier Enhanced Annotations, *Proceedings of the International Conference on Society for Information Technology and Teacher Education (SITE 2003)*, (2003) pp. 991-993. See also <http://dl.aace.org/11854>.
- [23] H. Krottmaier and H. Maurer Transclusions in the 21st Century, *Journal of Universal Computer Science*, 12 (2001), pp. 1125-1136. See also [http://www.jucs.org/jucs\\_7\\_12/transclusions\\_in\\_the\\_21st/](http://www.jucs.org/jucs_7_12/transclusions_in_the_21st/).
- [24] J. Lennon and H. Maurer Why it is Difficult to Introduce e-Learning into Schools and Some New Solutions, *Journal of Universal Computer Science*, 10 (2003), pp. 1244-1257. See also [http://www.jucs.org/jucs\\_9\\_10/why\\_it\\_is\\_difficult/](http://www.jucs.org/jucs_9_10/why_it_is_difficult/).
- [25] B. Leuf and W. Cunningham *The Wiki Way. Quick Collaboration on the Web*, Addison-Wesley, 2001.
- [26] C. C. Marshall Annotation: From Paper Books to Digital Library, *Proceedings of the ACM Conference on Digital Libraries*, (1997) Philadelphia, PA, U.S.A., pp. 131-140.
- [27] C. C. Marshall and S. Bly Sharing Encountered Information: Digital Libraries Get a Social Life, *Proceedings of the 4th ACM/IEEE-CS Joint Conference on Digital Libraries*, (2004) Tucson, AZ, U.S.A., pp. 218-227.
- [28] H. Maurer Beyond Classical Digital Libraries, *Proceedings of the NIT Conference, Global Digital Library Development in the New Millenium*, (2001) Beijing, China, pp. 165-173. See also [http://www.iicm.edu/iicm\\_papers/beyond\\_digital\\_libraries.doc](http://www.iicm.edu/iicm_papers/beyond_digital_libraries.doc).
- [29] H. Maurer and K. Tochtermann On a New Powerful Model for Knowledge Management, *Journal of Universal Computer Science*, 1 (2002), pp. 85-96. See also [http://www.jucs.org/jucs\\_8\\_1/on\\_a\\_new\\_powerful/](http://www.jucs.org/jucs_8_1/on_a_new_powerful/).
- [30] H. Mülner Software für ein multimediales Lexikon, (2001) Personal Communication.
- [31] T. H. Nelson *Literary Machines*, Mindful Press, 1981.
- [32] T. H. Nelson Generalized Links, Micropayment and Transcopyright, <http://www.almaden.ibm.com/almaden/npuc97/1996/tnelson.htm>, (1996) Visited May 1st, 2003.
- [33] K. Shafer et al. Introduction to Persistent Uniform Resource Locators, <http://purl.oclc.org/docs/inet96.html>, (2005) Visited February 24th, 2005.
- [34] E. Miller et al. Resource Description Framework (RDF), <http://www.w3.org/RDF/>, (2004) Visited February 3rd, 2005.
- [35] D. Libby, Netscape Communications RSS 0.91 Spec, revision 3, <http://my.netscape.com/publish/formats/rss-spec-0.91.html>, (1999) Visited February 3rd, 2005.
- [36] A. Swartz RDF Site Summary (RSS) 1.0, <http://web.resource.org/rss/1.0/>, (2001) Visited February 3rd, 2005.
- [37] J. Rubart et al. Organizing Shared Enterprise Workspaces Using Component-Based Cooperative Hypermedia, *Proceedings of the Twelfth ACM Conference on Hypertext and Hypermedia*, (2001) Aarhus, Denmark, pp. 73-82.
- [38] W. Treese Putting It Together. Open Systems for Collaboration, *netWorker*, 1 (2004), pp. 13-16.
- [39] F. B. Viégas et al. Studying cooperation and conflict between authors with history flow visualizations, *Proceedings of the 2004 Conference on Human Factors in Computing Systems (CHI 2004)*, (2004) Vienna, Austria, pp. 575-582.
- [40] A. Weiss Ending Spam's Free Ride, *netWorker*, 2 (2003), pp. 19-24.
- [41] Wikipedia. The Free Encyclopedia, <http://www.wikipedia.org/>.
- [42] WIKIPEDIA, *Wikipedia. The Free Encyclopedia*, <http://en.wikipedia.org/wiki/Wikipedia>, (2006) Visited July 6th, 2006.

Contact address:

Josef Kolbitsch  
Graz University of Technology  
Steyrergasse 30  
8010 Graz, Austria  
e-mail: josef.kolbitsch@tugraz.at

Hermann Maurer  
Institute for Information Systems and Computer Media  
Graz University of Technology  
Inffeldgasse 16c  
8010 Graz, Austria  
e-mail: hmaurer@iicm.edu

---

JOSEF KOLBITSCH is a PhD student at Graz University of Technology. His research interests include electronic encyclopaedias, digital libraries and hypermedia.

---



---

HERMANN MAURER is professor and dean of the faculty of computer science at Graz University of Technology. He is author of some twenty books and more than 600 contributions in various publications. Recently he has also published "XPERTS", a series of science fiction novels.

---