Proceedings of FREENIX Track:
2000 USENIX Annual Technical Conference

San Diego, California, USA, June 18–23, 2000

OPEN INFORMATION POOLS

Johan Pouwelse

Open Information Pools

Johan Pouwelse
Delft University of Technology, The Netherlands

j.a.pouwelse@its.tudelft.nl

Abstract

On the WWW it is not possible to supplement existing web pages of other people with new information or a link to that information, because the WWW does not have a standard method for write access. With write access, information can be added in the right context, which eases searching. We therefore define Open Information Pools: a collection of WWW based databases with public write access. By using databases we add structure to the information. Each database deals with a specific topic. We developed an architecture to support Open Information Pools. Important elements in the architecture are the rating and moderation tools. With these tools the user group is able to maintain and update the database and also to prevent errors and abuse.

We conducted measurements on operational rating and moderation tools to show the validity of our idea. The study of Slashdot.org's rating and moderation tools shows that insightful information is recognised after only 37 minutes. We implemented a prototype of a true Open Information Pool containing music information. This database contains biographies, audio CD descriptions, audio CD cover pictures, lyrics of the songs with timing information, and MIDI files. We developed several tools to create, insert and search this database.

1 Introduction

Several people have dreamed of building a system that could unlock the knowledge of humanity. The MEMEX system [1], Xanadu [17], and the Word Wide Web (WWW) are steps to realise that dream. Inspired by these ideas we propose a system that is one step further to the realisation of that dream.

The WWW contains vast amounts of information, without any structure. The lack of structure within WWW is both its strength and its weakness. Finding information in the vastness of the WWW is a serious problem. It also lacks some features for access and addition of information.

First, the WWW lacks the possibility of writing information in context. In the classic paper of Vannevar Bush [1] the addition of information to the collection was identified as a mandatory feature. However, WWW pages on the Internet do not have write access. Only a limited and non-standard form of write access can be provided through the use of CGI scripts. A contribution of information can *not* be made to the point at which it appears on the WWW. The inability of supplementing already present information with a bi-directional link inhibits the true accumulation of information. Information cannot be placed on the WWW within the context to which it belongs, instead it is tied to a “site-name”. As a consequence, search engines have to restore the context by giving the translation of keywords to relevant locations. Second, the WWW has no standard rating and moderation systems. Between the outdated, irrelevant, and incorrect content on the Internet lie the true gems of information. Without a built-in mechanism to rate and moderate information it is impossible to make a distinction. We need to learn from the mechanisms that already exist for a long time in the “paper world”. Third, there is no direct support for replication and synchronisation within WWW. A system that contains all knowledge of humanity needs to be distributed across the globe. Without any distribution the central points of the system would break down. Mirroring of sites and proxies are only added solutions to replicate the WWW. The WWW is not a fully distributed system, it is a collection of interconnected, yet independent HTTP daemons.

The WWW represents a significant advancement for the dissemination of knowledge, it can be viewed as a 15-billion-word encyclopedia [3, 12]. The significant growth in size, access, and reach of the WWW make it vital that new systems are enhancements of the WWW instead of competitive replacements. Addition of the above features to the WWW is difficult. Write access is a nightmare for security. Rating and moderation is difficult because ratings are highly individual and depend on a personal point of view. Distribution is an ambitious step compared to the current WWW practice.

The basis of our *open information pools* concept is our belief in openness. Information does not appear, but must be provided by someone. The obstacles for adding information must be as small as possible. The organisation of this information must be as open as possible: everybody must be able to contribute or copy content. With this openness we are able to solve the three problems of the WWW.

This paper applies the ideas of Open Source software to collections of structured information. The Open Source idea is that software must be distributed as source code, enabling everybody to improve and extend this software [9]. Due to a special clause in the GPL copyright notice, often used by much of the Open Source Software community, software under a GPL copyright must remain in the public domain. The philosophy of keeping something open has been applied to other fields besides software as well. For example, an initiative has been started to keep web content open [25], and another to publish hardware designs [29].

This paper proposes to create open information pools, with the underlying philosophy that *information needs to be in the public domain*. We define an open information pool as "*a structured collection of text, pictures, movies, sound, and other data, which may be freely copied and to which the whole Internet community can add information*". The structured pools of information are implemented as relational databases. Each database contains information about a specific topic in several tables. Everybody can contribute to such a database and also copy the content, provided that added information remains public. Rating and moderation systems are used to keep the content "clean". The open form makes it possible to create information pools that are impossible, or not cost effective, to create and maintain by a firm. When a firm owns and maintains a database, the database can dissappear when for example the firm goes out of business. Public information has a smaller risk of disappearance.

In [9] Raymond discusses the gift culture of Open Source. With our open information pools we will try to show that the same gift culture can also be cultivated in the information domain, besides the software domain. Within Open Source it has been shown that people are willing to freely share their work without any direct payment. The quality of this shared work can even surpass the quality of the work produced within firms that do not use this gift culture. On the Internet people are willing to share information freely, for example, within Usenet. People can be motivated to share information for free for the satisfaction of a good reputation amongst their peers, the knowledge that a good reputation can also pay-off in the real world, or pure altruism.

By using the Internet as the basis of the information pools, world wide access is guaranteed. Adding information to Internet-based databases is already possible [20, 21, 23, 24, 25, 26, 27, 28]. The majority of these databases do not contain scientific, medical, historical, or cultural knowledge but "popular information", with a wide audience. Unfortunately, several of these databases have been taken out of the public domain for commercial reasons. Users cannot copy the entire content of those non-public domain databases. We believe that this protective commercialism is a dangerous development that threatens the free flow of information. The Internet movie database [28] (about 200,000 entries), and the audio CD database [21] (about 500,000 entries) were open, but are now closed. The audio CD database is popular, with more than 600 contributions of information *daily*. However, the quality of the content is low; the number of duplicate records is measured to be as high as 12 in some cases. Slashdot.org, in contrast, is a good example of how an open information pool should operate. Slashdot.org is a web site with an open news and discussion forum, it has public write access, is based on a relation database, uses an advanced moderation system, and has a large user base. The Open Directory Project [26] is also a good example, they have created a comprehensive directory of over a million web pages, by relying on a "vast army" of volunteer editors. The directory can be freely copied, and is used by Netscape, Lycos, HotBot, and others.

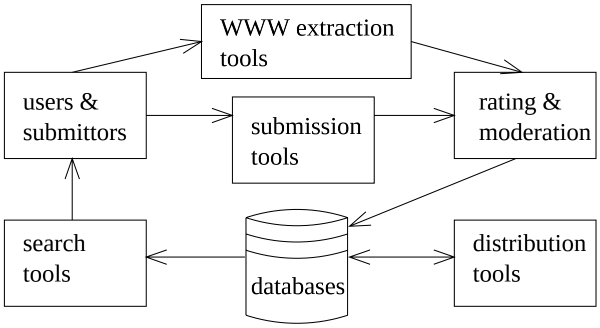


Figure 1: Open Information Pools architecture

2 System overview

We designed an architecture that supports open information pools and adds to the WWW the three missing features discussed in Section 1. The architecture consists of several components, see Figure 1. Current Internet based writable databases are tightly bound to the database content. In our architecture, every component is independent of the information pools content. With our generic architecture, the number of these open, Internet based information pools could grow and software could be re-used.

2.1 Users and submitters

The submitters are the key to the whole open pool concept. The database must be of sufficient interest in order for them to freely submit the information they have. People that have bought a described item, are fan of the subject, or do research on it are especially motivated to submit. Computer literacy also plays a role, computer related database entries such as security vulnerabilities, $2nd$ hand computer adds, FAQs, manuals, and computer science encyclopedia could be popular. When information is submitted it is also important to register the submitter as a form of recognition for his efforts. People with a high number of submissions build a certain reputation with this work. The motivations behind submission are thus not limited to pure altruism. The influence of reputations in a gift culture are explored by Eric S. Raymond in [9].

2.2 Submissions, rating, and moderation

The power of the open information pool is also its greatest weakness; the open form makes the information pool vulnerable to errors and abuse. The rating and moderation component ensures that the database is kept clean. Users can submit information into the pool with submission tools. Submissions can be of two types: inserts and modifications. For the storage of modifications within an information pool there are two methods: it is possible to store the original database and the subsequent modifications, or to store only the latest updated version. What method should be used depends on the need for history within the information pool. Different rating and moderation policies can be used for inserts and modifications. For example, an information pool in the form of an encyclopedia for computer science can have the following policy: every Internet user can freely insert new terms and descriptions, and a team of moderators is appointed to modify and delete entries. With such an open policy of an information pool, attacks on integrity and availability cannot be prevented. Write access could be blocked for new users, users only obtain write access after using the information pool for some time. Some boundaries could be set on number of inserts and modifications. When anonymous inserts are not allowed, limits on the number of submissions per user can be set. Another possibility is to limit the number of these transactions per computer by using the IP address. Numerous, slightly different policies can be constructed around an identical central mechanism, several policies can also be mixed to form a new, hybrid policy.

2.2.1 Submitter based

A simple policy is to place the first submitter of the content in control of subsequent modifications. New information may be freely inserted. With every insert the e-mail address is requested. Modifications are allowed but have to be approved by the original submitter, using e-mail. If the original submitter does not respond within a reasonable amount of time, he loses his approval rights. The original submitter is then automatically replaced by the submitter of the modification. This policy is the most simple to use and implement, but it does not offer strong protection against errors.

2.2.2 Reputation based

A more advanced system is build around the reputation of users. Each user is registered with a (nick)name and a number that indicates his reputation. A reputation function calculates a reputation based on several aspects of the user, like the amount of activity of the user, number of retrievals of his submitted information, the number of modifications to his submissions, etc. Based on this reputation all sorts of moderation methods become possible. For example, a high reputation enables people to moderate content submitted by people of significantly lower reputation. This policy requires registration and identification of the user with for example an account and password. This gives a barrier for submission of information, but also more protection. When a new account is requested the password is e-mailed to the user, thus a valid e-mail adress is required. A reputation-based system is more complex, provides a small barrier, and offers stronger protection against errors than the submitter based policy.

2.2.3 Democratic based

With a rating mechanism the majority determines what is considered "correct". Users indicate the quality of every insert and modification. Quality ratings can consist of numbers that indicate the level of accuracy, completeness, or grammatical correctness, it is also possible to rate quality with a "better/worse than" indication. It is very difficult to rate the content of a *web page* because the rating is strongly dependent of the point of view and the purpose of the text, according to [10]. Because a database has more structure and a high level of context, we belief accurate quality rating is possible. When users search through the information pool the ratings can be used to filter out information below the desired quality level. The democratic and reputation based policy can be combined into a system where votes of users are weighted with their reputation, and (dis)agreements influence the reputation of the submitter. A problem with the democratic policy is that users cannot always be objective and independant, for example the majority can determine that false popular belief is true.

2.2.4 Expert based

Experts are (democratically) appointed users that control all the inserts and modifications. Users can freely submit inserts and modification, but the experts determine if they should be added to the open information pool. A variation of this policy is to create a hierarchy of experts. Using the expert policy places a very large responsibility at the experts. Problems may arise if experts do not have sufficient time for moderation, lack of interest for their task, or if their objectivity is questionable. This policy is used within the Open Directory Project [26].

2.3 Databases

A database can contain text, pictures, music, movies, and other data. Examples of possible databases include a history database with people, places and events, encyclopedia on various topics, tutorials on various levels, consumer reviews on products, medical information, product pricing per (e-)store, stock market numbers, TV listings, etc.

Representation of information is a difficult subject. The information stored by the WWW is not structured, hence the context of information is hard to find. Recent WWW developments include adding meta-tags or database extensions that describe the context of WWW pages [6]. When a database is used for storing the information the context is well defined and the content is highly structured [19]. Using a relational database for the storage of web-based information is more powerful than the method of using standard files. The additional power of relational databases is particularly strong when using large pools of structured information.

The demands and properties of a database depend on the class of information stored. The first separation we make is between objective and subjective information. The former is an unbiased truth about subjects, phenomena, people, places, objects, companies, etc. The latter is a person's belief or opinion about some subject, people, etc. with a biased truth. The second separation we make is between deterministic information and non-deterministic information. When there is only one answer and one logical textual representation to a question we call it deterministic information. For example the question "In what year was Napoleon Bonaparte born?" results in a single answer. The question "What is the life story of Napoleon Bonaparte?", can be a complete paragraph or book of non-deterministic information.

Objective information is more easy to store and to maintain in the database than subjective information. Moderation is simplified when the content is free of beliefs and opinions. When non-deterministic information is stored in the database, the demands on the rating and moderation tools are higher. Non-deterministic information is very hard to capture and maintain in a value of a field in a database. The reason for this is that modifications on non-deterministic information in text form can be replacements of the whole text, small modifications on several sentences, a new combination of several other suggested modifications, a new structure of the whole text without modification of sentences, etc.

2.4 WWW extraction tools

Open information pools are not replacements of current WWW pages. With WWW extraction tools we can add context and structure to the WWW content and insert it into an open information pool. Extraction tools are very useful to start a new information pool and re-use WWW based content. Several extraction tools exist that can extract the information from a WWW site [15, 18]. Most extractions tools use a *wrapper* for each WWW site. A *wrapper* is a program that extracts information from a specific WWW page and presents this information to the user or inserts it into a database. Because each WWW site is different, wrappers are unique to each WWW site. The extraction tools can generate these wrappers after they are configured for a particular WWW site by the user. This user configuration is time consuming. In [5] software is described that can detect the structure of a web page automatically. This software does not require user configuration, but the software is reported to show "meaningful" results in only 70 % of the cases. Even with these advanced WWW extraction tools, submitters must find WWW sites that contain information missing in the information pools. When automatic WWW extraction tools are used for submission, the submitter information must be present to enable moderation. The extraction tools must therefore be configured with information about the submitter. This, however is not necessary when anonymous submissions are allowed.

2.5 Search tools

Search tools can help the user to search through the information pools. Search tools are more effective than the general Internet search engines because the information is stored in a more structured way, within a known context. In [13] a distinction is made between *discovery queries,* used by WWW search engines such as Yahoo, Altavista, etc. and *retrieval queries* that are queries on a specific collection of WWW pages (intranet) that are well maintained, and have a known structure. WWW search engines often fail to exploit the structure of such WWW pages, because the exact semantics of links remain invisible. Because WWW search engines cannot exploit structure, they prefer indexing "flat" WWW pages.

Discovery queries of Internet search engines ask keywords to make a sorted list of relevant WWW pages. Often these discovery queries result in a large numbers of matching pages. The coverage and recency of Internet search engines (Altavista, Excite, HotBot, Infoseek, Lycos, Nothern Light) is extensively analysed in [12]. This study, published in 1998, estimated that the lower bound for the number of WWW pages is 320 million. It is also shown that no single Internet search engine indexes more than one-third of the "indexable WWW". In contrast, retrieval queries on information pools do not suffer from the effect that information is not "indexable". A search is broken into two steps. First, the appropriate information pool is located. Second, this single database is searched to answer the query.

2.6 Distribution tools

For performance and reliability the information pools can be replicated. The openness of the information pools also dictates that there must be no single person or organisation in control. Information must be freely shared among all interested users on the Internet. Users must be able to download a whole database and access it off-line. Updates must be exchanged between different sites to keep them synchronised. Updates must not be controlled by a single site in a scalable distributed system. There are several update mechanisms. (1) Updates can be distributed in the form of a chain where each location forward the received updates and his own updates to the next location. (2) A hierarchical system can be used distributes the updates from a group of top level systems. (3) The use of a news group to broadcast the updates. A news group with formatted e-mails is a particular useful method of distributing updates because it builds on a world-wide broadcast mechanism. The reliability of a system using news groups is higher than the alternatives, yet more bandwidth and other resources are used.

A nice property of our information pools is that they are self-describing. Having a description of the databases off all information pools is very usefull. A special database called *meta* is used to describe the various databases, the policy they use, the tables and fields therein, the distribution method, and the locations were they are stored. The open meta database has a special moderation method. The inserts can only be made from the site which hosts the open database. This can be simply checked using the IP number. The meta database is useful for search tools because it contains the context of all information.

3 Legal issues

Copyright laws on the Internet have always been under pressure because of the ease of duplication. Redistribution and publication of copyrighted material is in most cases not allowed. Copyrighted material *cannot* be entered into an open information pool without permission of the copyright holder. For a good introduction about copyright laws and databases, see [2, 14]. It is impossible to claim the copyright of facts, ideas, and public information. The non-trivial question is, what material is copyrighted, and what is an unprotected fact, idea, or public information.

The best legal term describing our information pool concept is the term "automated database". The copyright law of the United Sates defines an automated database as: "*a body of facts, data, or other information assembled into an organised format suitable for use in a computer and comprising one or more files*" [11]. The copyright laws of the US and many other countries only protect "original work". The definition of an original work was traditionally coupled to some form of *creativity* within this work. However, it is very hard to show the creativity for the content of an automated database. To extend the copyright protection to factual automated database the "sweat of the brow" doctrine was introduced. But this doctrine was recently abandoned in a US supreme court ruling of *Rural telephone Service Co., Inc v. Feist Publications, Inc.* 663 F. Supp. 214 218 (1987). Open information pools of facts and public information can therefore not be copyrighted because there is no original work, no creativity, and the "sweat of the brow" doctrine was abandoned. The implications for the open information pool concept are that making a copy of an existing database on the WWW with facts, ideas, and public information is allowed, but copyrighted information remains off-limits.

4 Implementation

In this section we first discuss measurements we conducted on rating and moderation tools. Second, we discuss our own prototype that is the first step towards a full implementation of our open information pool concept.

4.1 Rating and moderation measurements

Rating and moderation tools have been discussed extensively in the literature, see for a good example [10], and for an overview [4]. However, we were unable to find any publication on the performance of such tools with both *real* content and a large group of *real* users. To investigate the dynamics of rating and moderation tools, and to support the validity of our open information pool concept we conducted measurements on a web site with operational rating and moderation tools. We choose the Slashdot.org web site because it has already at lot of the elements we use in our open information pool concept. Slashdot.org shows important news items on various topics in computer science and gives readers the ability to freely attach *news comments* with remarks, enhancements, and insights to a news item. Inserted news comments can be viewed by all readers and are attached to the original news item. All news items and comments are stored in a large relational database. Slashdot.org uses a hybrid rating and moderation system that is based on the reputation and democratic mechanisms discussed in Section 2.2. However, Slashdot.org is *not* generic, it is fully dedicated to the "news forum" purpose. The

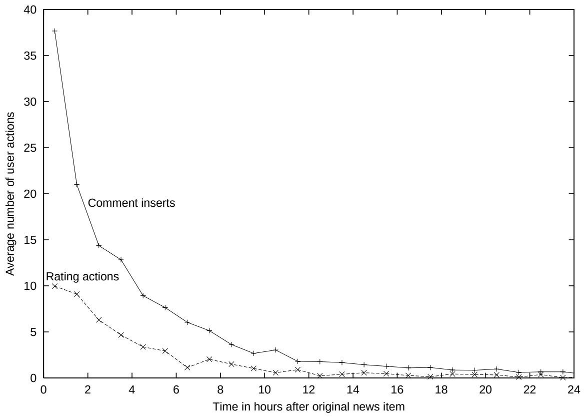


Figure 2: Slashdot.org user actions.

database routines, rating system, and moderation system cannot be re-used for other classes of information besides news items. Another important limitation is that news comments cannot be modified once inserted. The lack of modification options is not important for news comments, but does prevent news comments from being corrected for spelling errors. A generic implementation of an open information pool must allow several operations on the database content to update or improve it. When the content is text, the operations are for example correcting spelling errors, joining several paragraphs together, making modifications on several sentences, etc.

Within Slashdot.org every comment to a news item is rated with a *score* from -1 (inferior) to 5 (insightful) by the readers. The initial score of a news comment is 1 by default, 2 for users with a high reputation, and 0 for anonymous news comments. A percentage of the (randomly-selected) readers with sufficiently high reputation are allowed to influence the scores. When users are asked to influence the score they can add or subtract a single point to the score of a news comment. The reputation of a user is calculated based on the score of his recently posted news comments. Readers can set the level of moderation by filtering news comments below a score threshold. If the lowest threshold of -1 is selected, all news comments are visible.

Comment score	Percentage
inferiour, -1	8.1 %
0	31.0 %
1	40.3 %
2	14.5 %
3	3.8 %
4	1.2 %
insightfull, 5	1.1 %

Table 1: Slashdot.org comment score distribution.

In Figure 2 two user actions are shown: the number of new added news comments and the number of rating actions. The news comments are the average number of inserted comments per hour in the discussion of a single news item. The time is set relative to the publication of the news item. Each rating action is the addition or subtraction of a point from the news comment score. The results are calculated from an analysis of 30 news items which received more then 4,250 comments and were subject to 1,400 rating actions. The average number of inserted comments per news item is about 142 (4250/30). This figure shows that the activity of the users is fairly high and fast, on average half of the comments is inserted within three hours of the news item's release. Within the first hour of the news item release almost 38 news comments are inserted. After this initial attention the user activity drops fast and after 12 hours less than two comments are inserted per hour. The rating system does not show this sharp decrease. On average almost 10 rating actions are performed within the first hour on the 38 inserted news comments. After 10 hours, less than one rating action is carried out per hour.

The resulting distribution of the news comments scores is shown in Table 1. If a reader selects the threshold of 3 for reading, 93.9 % of the news comments are not shown. Unfortunately it is very difficult to determine objectively if the remaining high ranking comments are really of high quality. It is the personal opinion of the author that the score of a news comment gives a good indication of the quality of the comment. Deviations between the score and the quality are not frequent and are seldom more than a single point. Overall the rating and moderation system works effectively. For example, news comments containing strong language with no insight, no intelligent remarks, and no enhancement are quickly rated -1, (inferior).

The required time for a news comment to obtain the final score is shown in Figure 3, for clarity only the four major start and end score combinations are shown. Each change in the score is a result of a user rating action. Time is taken relative to the insertion of the news comment. The number of measured news comments (n) with this start and end score

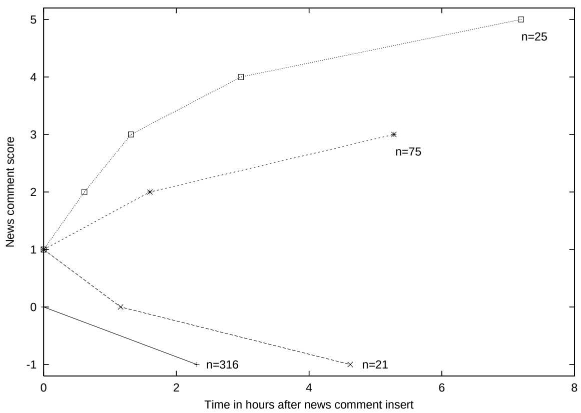


Figure 3: Speed of the rating system.

combination is shown at the endpoint of the lines. It is important to know that the initial score is changed for only 21.8 % of the inserted news comments. On average it takes 37 minutes for news comments with an initial score of 1 and final score of 5 to receive the first addition to the score. For news comments with an final score of 3 the first addition comes after 1 hour and 36 minutes. Receiving the last point towards the final score takes significantly longer. For the following [initial, final] score combinations [1,5], [1,3], [1,-1], [0,-1] the last rating point is received after 253, 220, 207, and 138 minutes respectively. This indicates that the convergance towards the final score starts fast and than slows down.

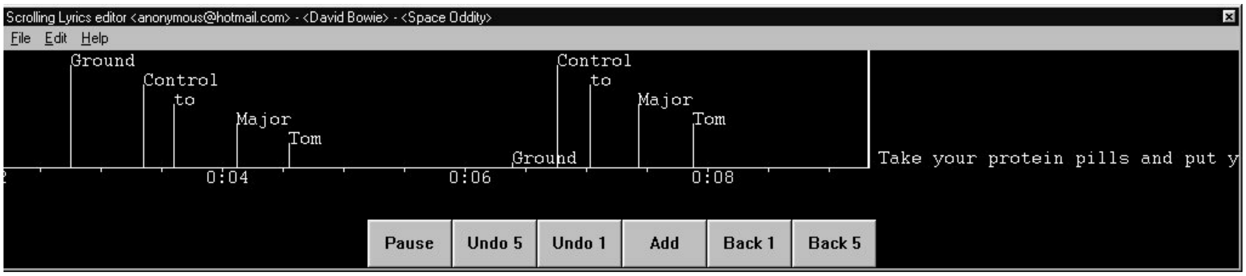
To summarise we note that the rating of the news comments by the users works effectively. The users are very active with both inserting comments and rating comments. The rating system of Slashdot.org works on a surprisingly small time scale; insightful news comments in the database get their first reward after only 37 minutes. With the measurements on Slashdot.org we see that subjective, nondeterministic information (see Section 2.3) can be rated fast and accurately. Another important characteristic is that users are free to determine the level of moderation by setting a filtering threshold.

4.2 Prototype

We are currently building a prototype of a system for a single open information pool, with all the elements of the overall structure outlined in Figure 1. Our first goal is to develop new rating and moderation tools from studies with actual users and real content that work effectively for various classes of information and allow all sorts of modifications. The next step is to build a generic implementation that enables a user to create a new open information pool using only a WWW interface. With this WWW interface the user selects a rating method, moderation method, tables, fields, and other options. The system would automatically create the information pool and insert this new pool into the meta-database (see Section 2.6) and allow all Internet users to browse and submit information.

For the database content of the prototype we have chosen music information. The database consists of music artists/band biographies, produced audio CDs, audio CD cover pictures, songs on every audio CD, lyrics of the songs with timing information (≈karaoke), and MIDI files. There are four reasons for this decision. First, this content has a high degree of context, requires frequent updates, and combines various information classes (see Section 2.3). This is important because we want to study rating and moderation systems for all information classes in action. Second, the content of this database is attractive for a very large and active user group. Third, because the existing database with some of this information called "Audio CD database" [21] is taken from the public domain, see Section 1, we want to return this database to the public domain. Fourth, we can re-use and extend the source code of the CDIndex project [20]. The open source CDIndex project is developing software for an *open* audio CD database as an alternative for the *closed* audio CD database. The aim of CDIndex is the same as the original audio CD database, focused to storing audio CD "table of contents". The table of content of an audio CD consists of the names of the songs and artists, this information is not present on the CD itself. CDIndex has an operational database and produced working code to anonymously insert information. Rating, moderation, and distribution tools are not yet developed and the database is virtually empty.

Our prototype is implemented in Perl on a Linux system using the MySQL database and Apache WWW server. The music information database of the prototype is finished and can be viewed, searched or filled from an ordinary browser. Submission, voting, and moderation tools are still under development. We implemented a generic search tool that can be used for keyword searches and to browse through a database. This generic search tool queries the relational database for tables and fields that can be searched. This information is not present



Scrolling Lyrics editor <anonymous@hotmail.com> - <David Bowie> - <Space Oddity>

Take your protein pills and put y

Figure 4: The Scrolling Lyrics creation and submission tool.

in the generic search tool. The search tool uses the database structure information to show a HTML search form. A basic WWW extraction tool has been made. This WWW extraction tool was configured to extract audio CD cover pictures, MIDI files and lyrics from several WWW sites. The open information pool now contains more than 10,000 audio CD covers, numerous MIDI files, and an extensive amount of lyrics.

A specific client program called “Scrolling Lyrics” is developed for the display, creation, and submission of the lyrics with timing information. Scrolling Lyrics is an Open Source plug-in for the popular Windows95 music player called “WinAmp”. A screen shot of Scrolling Lyrics program in edit mode is shown in Figure 4. A user can enter the lyrics, add a timestamp for every word in the song, and submit this information to a server of their choosing. In the screen shot, text with time stamps is shown as well as plain text. When the music is playing users can click the “add” button to add a timestamp to a word in the song. The text with timestamps has a staircase effect for readability. When the scrolling lyrics program is in play mode, the lyrics of the song scroll in synchronisation with the music. Options are included to save this scrolling lyrics information inside a compressed MPEG audio file (mp3) and to submit it for entry in the music database. Submissions are formatted in XML, a more general form of HTML. The submissions are delivered to a server with the HTTP Post method. The processing is done with perl CGI scripts, they insert the submission in the music database. The distribution tools can optionally post every submission that arrives at the server directly in a news group.

There are several unsolved legal issues tied to this music database. The lyrics of almost all popular songs are copyrighted and cannot be entered into the music database without permission. If the scrolling lyrics information is a “new and original work” it is not copyrighted and can be entered freely. The question of copyright on scrolling lyrics information is unresolved. Under the local Dutch copyright legislation it is permitted to give people a personal copy of lyrics without violating the copyright. Using scanned images of audio CD covers is also permitted, provided that they are used for on-line sales catalogs, for example an electronic audio CD shop with a preview image of the audio CDs on sale.

The source code of our prototype and the scrolling lyrics program is freely available from [22] under the Open Source license. The measurement software for the Slashdot.org rating and moderation tools is available on request. This measurement software can interfere with the proper operation of the Slashdot.org server.

5 Related work

We have not been able to find any earlier work raising our basic idea - a world-wide collection of open, public-writable databases with moderation and world-wide distribution. However our idea touches on a number of other publications in different parts of computer science, including hypertext, databases, knowledge bases, etc.

Various researchers have investigated the combination of relational databases and the WWW. In [6] the idea of including the database tables, attributes, and relations inside HTML was presented. The complete separation of content and appearance is argued in [19], where the content is organised as a database. The developers of RCS and CVS [8] have created tools to update, synchronise, and distribute files. These tools provide some basic support for public write access and moderation of submissions coordinated by a central server. CVS operates on the file and directory level and provides an abstraction level that is very basic, for example database records are outside the scope of CVS. Our concept is different from CVS because we use a distributed structure, operate on the database level, and use generic rating and moderation tools. As early as 1990 the first prototype browser appeared with both read and write capability, yet the WWW has primary been used as a read-only tool. The recent WebDAV standard [16] transforms the read-only WWW into a writable, collaborative medium. Access is restricted and cryptographic authentication schemes are used to enforce this. WebDAV defines mechanisms for file locking, version management, hierarchical organisation, and access control. The standard does not specify distribution, rating, and moderation tools. It differs from our approach because it is designed only for use by a closed group of users and cannot be applied on a global Internet scale. Annotation and rating of WWW pages has been explored in [4, 10]. The annotation tools proposed in these papers use separate annotation servers and are limited to adding comment to existing WWW pages. This differs from our approach, we define integrated rating and moderation tools for structured information. The controversial US Communications Decency Act [7] increased the interest in such content rating and moderation systems. The Usenet system has some similarities to our concept. Usenet is also fully distributed and has public write access, but it does not have standard rating tools and enforces no rigid structure. The hierarchy of news groups with different subject areas is very loose when compared to the rigid structure of a relational database.

6 Conclusions and future work

We have introduced the idea of structured information pools that are open to the public. By giving public write access, information pools grow with each submission. We are working on a generic open information pool implementation based on a web server and relational database system. With this implementation a number of WWW drawbacks would be removed. Generic rating, moderation, distribution, and WWW extraction tools are needed to fill these open information pools and to keep them clean, accessible, and reliable. With our system we want to counter the trend that information is taken from the public domain. In the near future we will conduct a long term study to determine the dynamics of a large scale and intensively used open information pool to refine our implementation. We hope that our next implementation will be another step in the realisation of the dream that all knowledge of humanity will be unlocked.

Acknowledgements

The author wishes to thank Clem Cole and Koen Langendoen for their valuable advice, and the anonymous reviewers for their good suggestions.

References

[1] Bush V., “As We May Think”, Atlantic Monthly, July, 1945.

[2] Barlow J.P., "The Economy of Ideas: A framework for rethinking patents and copyrights in the Digital Age", wired, March 1994.

[3] Barrie J., Presti D., “World Wide Web as an instructional tool”, Science, 274, 371-372, October 18, 1996.

[4] Bouvin N.O., “Unifying strategies for web augmentation”, ACM Hypertext ’99 conference Doctoral Consortium, February 1999.

[5] Cohen W., ”Recognizing structure in web pages using similarity queries”, Proceedings of the Sixteenth National Conference on Artificial Intelligence, 1999.

[6] Dobson S.A., “Lightweight databases”, 1st Workshop on Logic Programming Tools for INTERNET, September 1996.

[7] Exon J., “Communication Decency Amendment”, US Senate, July 1995.

[8] Grune D., “Concurrent Versions System, a method for independent cooperation”, technical report 113, Vrije Universiteit, Amsterdam, 1986.

[9] Raymond E.S., “Cathedral & the Bazaar”, Sebastopol California, O’Reilly & Associates Inc, ISBN 1-56592-724-9, www.tuxedo.org/~esr/writings.

[10] Roscheisen M., Winograd T., Paepcke A., “Content rating and other third-party value-added information defining and enabling platform”, CNRI Journal D-Lib, August 1995.

[11] United States Copyright Office, “Copyright Registration for Automated Databases”, publication Circular 65.

[12] Lawrence S., Giles L., “Searching the World Wide Web”, Science, Vol. 280, Num. 5360, page 98, 1998.

[13] Lim E.-P., et.al., “Querying structured web resources”, Proceedings of the 3rd ACM International Conference on Digital Libraries, Pittsburgh, Pennsylvania, June 1998.

[14] Losey R.C., “Practical and legal protection of computer databases”, Orlando, Florida, 1995, FloridaLawFirm.com/article.html.

[15] Mecca G., et.al., ”The Araneus web-based management system”, in exhibits program of ACM SIGMOD, 1998.

[16] Whitehead E.J., Wiggins M., “WebDAV: IETF standard for collaborative authoring on the web”, IEEE Internet Computing, page 34 - 40, September - October 1998.

[17] Nelson T., “Xanalogical Media: needed now more than ever”, *to appear in* ACM computing surveys, hypertext issue.

[18] Sahuguet A., Azavant F., “W4F: a WysiWyg Web Wrapper Factory”, Technical Report, University of Pennsylvania, 1998.

[19] Sandewall E., “Towards a world-wide database”, 5th International WWW conference, May 1996.

[20] The open CD database, www.CDIndex.org

[21] The audio CD database, www.cddb.com

[22] The ultimate music database, www.mp3.nl

[23] Common vulnerabilities and exposures list, cve.mitre.org

[24] Archarology index, www.openarchaeology.org

[25] Open content initiative, www.OpenContent.org

[26] Open directory project, www.dmoz.org

[27] Link everything on-line, www.leo.org

[28] Internet movie database, www.imdb.com

[29] Open hardware design, www.lart.tudelft.nl