Int J Digit Libr (2013) 13:65–75
DOI 10.1007/s00799-012-0097-9

Can the Web turn into a digital library?

Hermann Maurer · Heimo Mueller

Received: 7 May 2011 / Revised: 19 August 2012 / Accepted: 21 August 2012 / Published online: 12 September 2012

Abstract There is no doubt that the enormous amounts of information on the WWW are influencing how we work, live, learn and think. However, information on the WWW is in general too chaotic, not reliable enough and specific material often too difficult to locate that it cannot be considered a serious digital library. In this paper we concentrate on the question how we can retrieve reliable information from the Web, a task that is fraught with problems, but essential if the WWW is supposed to be used as serious digital library. It turns out that the use of search engines has many dangers. We will point out some of the possible ways how those dangers can be reduced and how dangerous traps can be avoided. Another approach to find useful information on the Web is to use "classical" resources of information like specialized dictionaries, lexica or encyclopaedias in electronic form, such as the Britannica. Although it seemed for a while that such resources might more or less disappear from the Web due to attempts such as Wikipedia, some to the classical encyclopaedias and specialized offerings have picked up steam again and should not be ignored. They do sometimes suffer from what we will call the "wishy-washy" syndrome explained in this paper. It is interesting to note that Wikipedia which is also larger than all other encyclopaedias (at least the English version) is less afflicted by this syndrome, yet has some other serious drawbacks. We discuss how those could be avoided and present a system that is halfway between prototype and production system that does take care of many of the aforementioned problems and hence may be a model for further undertakings in turning (part of) the Web into a useable digital library.

Keywords Digital libary · WWW library · Wikipedia · Austria-Forum · Information consolidation

H. Maurer (✉)
Institute for Information Systems and Computer Media,
Graz University of Technology, Graz, Austria
e-mail: hmaurer@iicm.edu

H. Mueller
Institute for Pathology, Medical University of Graz, Graz, Austria
e-mail: heimo.mueller@mac.com

1 Introduction

The Web has turned into a dominant source of information. Most persons use it by employing one of the available search engines, or by going directly to a site they tend to rely on. Using one of the major search engines like Google.com or Bing.com is tempting, yet one has to be aware of a number of problems: first, often hundreds of thousands hits are presented, more than anyone will ever look at; actually, even if the search engine indicates that it has found some hundreds of thousand entries only the first few thousands can be accessed by the user: this goes little noticed since even reading beyond the first few pages of search results is more rare than it should be; second, most current search engines are still based on a set of words, rather than understanding natural language questions; third, the ranking of search results is not transparent; indeed it often depends on factors that influence the user in the wrong direction; (Note in passing that those three problem do not occur to any extent when using the "computational" search engine wolframalpha.com); fourth, the reliability of information found is not at all guaranteed: it is up to the user to investigate whether results can be trusted or not and this is often almost impossible. We will examine those four points in Sect. 2 with emphasis on the behaviour of Google and Bing.

Rather than using a search engine one might directly go to a specialized site. In Sect. 3 we discuss pros and cons of going to one of the sites that are structured in classical fashion, like a specialized collection of information on art, or animals, or

minerals, etc. We exclude Wikipedia from this set on purpose since we will discuss issues concerning Wikipedia separately in Sect. 4. In Sect. 5 we present a system, where an attempt is made to eliminate most of the weaknesses discussed in previous sections and argue that using the ideas mentioned we could finally end up with a very large repository of reliable material that we can base our work and our judgements on. We will also mention a new kind of E-Book collection that is part of our prototype that may go a long way to turning large portions of the Web into a genuine digital library: rather than treating books as stand-alone volumes with no connection with other material (beyond searching) we are in the process of testing algorithms that link information on book-pages to relevant pages in the same or other books, or to pages in our Wiki-system, and conversely. How to accomplish this is on-going research but we will present some concrete examples.

2 Some aspects of search engines

One of the most obvious problems encountered when using search engines is that the number of search results is too large to be used systematically. Further, many search results are similar to others, i.e. have an undesirably high degree of redundancy; worse, some search results are contradictory.

Concerning redundancy, it would be nice if search engines in the future would try to cluster together similar search results automatically, may be even combining results in a cluster into one more or less coherent document, so that users are only confronted with a limited number of clusters, or even better with documents representing the most important "views" on a topic.

In some isolated cases, this has worked quite well as pointed out in [31]: it is shown in that paper that in some cases redundancy can be cut by 75 % applying fairly simple similarity recognition algorithms like in [21] or in [29] or algorithms used for plagiarism detection, like in [9,19,20,32]; and [14]. To be more specific, 20 essays on 50 topics were reduced to an average of 6.3 essays per topic without loss of information. The only price paid was that not all essays were as coherent and smooth to read as the originals.

To reduce redundancy dramatically (not by 75 % but by 99 %) and to retain coherent essays (i.e. to construct coherent essays from lots of snippets that have been collected) is still something that sounds like science fiction today; yet it is one of the greatest challenges search engines are facing, even if finding solutions will still take some major breakthroughs in the semantic analysis of textual corpora. Yet clustering similar documents together (using word-vector approaches or such) and automatically preparing a few sentence summary of the difference of one cluster when compared with a different cluster is possible today.

Using good visualisation techniques, the relation between document clusters could be shown in an impressive way: it is a pity that the publicly available large search engines are not making use of those features to any extent. Again, it should be noted that there is continued progress: during the time of the submission of this paper and its appearance Google Knowledge Graph has made its entrance: using it, when locating a document, related ones are shown. Google is able to do this using rich information from previous queries but misses out on one significant aspect: the edges in the knowledge graph have no meta data associated with them; thus it is not immediately visible whether one node is supporting, contradicting or answering a question, etc. At the moment, search engines are reducing the amount of information available to users mainly by ranking and by "personalizing" them. They should also allow to do further searches in the large set of documents located (something that would help a lot and would be easy to offer) or to allow the user to first narrow down their area of interest, with first attempts visible in search engines such as the slash-tags in [3]. With Blekko you can narrow down your search by prefixing it with a series of "slash-tags". Thus, e.g., Iceland volcano /flickr gives you what you would expect: pictures of volcanoes in Iceland from flickr.

A major problem is the issue of ranking. Ranking algorithms are usually not publicized, giving rise to many speculations. Like, if an item A is listed before an item B, is it really because it is the better hit, or is it that whoever is behind A has a better relationship with the search engine, maybe even to the extent of paying for preferred treatment? Is the sometimes spread rumour that sites using Google Analytics are on purpose better indexed by Google as bonus for using another Google product true or not?

It is probably of interest to some readers to note how ranking can have a negative effect for them. Let us explain this in terms of a (very realistic) example. When trying to book a hotel in some city one often finds easily some booking agency that allows you to book a suitable hotel at ease. Although all kinds of information on the hotel is provided, like how to get there, amenities available etc., a valid phone number or E-mail of the hotel is often not included for obvious reasons: one does not want the customer to be able to make direct contact with the hotel.

Although this is usually not of concern, it may be, if a last minute change becomes necessary, if one is lost and wants to phone for directions, if one has to cancel or re-book, etc. Seasoned travellers overcome this barrier with a trick: they use the booking agency to locate a hotel that looks good; then they use the hotel's name for a search with a search engine. This will usually lead them again to a or the same booking agency, but now comes the trick: going some ten pages further in the search results gives a good chance to find the website owned by the hotel, with all necessary parameters. With some luck, one can even get rooms cheaper,

one can certainly negotiate for some small extras that one usually would not be able to mention. ("My room should not be adjacent to the elevator shaft" might be a valid concern by persons who are disturbed in their sleep because of the audible movement of the elevator.) It is interesting (disturbing?) to notice that some agencies are starting to refuse to accept hotels for bookings if the hotels have their own homepage! Concerning correctness, let us quote from [31]:

> "We all accept that no information obtained is reliable (except if know we can trust the source of information), yet how dramatic the unreliability is can be shown with numerous examples. Searching for "boiling point of Radium" with Google two entries retrieved Aug. 25, 2010 are shown in Fig. 1.
> One entry shows 1737 degree Centigrades, the other 1140. How should we know which one is correct?
> May be life does not depend on this particular answer. However, consider a case we have been confronted with when we picked a type of wild mushrooms recently that we could definitely identify as "Echter Ritterling" (Gruenling). When we wanted to check if it was edible or not we found five entries on the first search page, three telling us that it is a delicate edible mushroom, one informing us that it is deadly poisonous and one simply that it is poisonous!"

How is it possible that even in what seems reliable sources such wild discrepancies and contradictions occur? There are two main reasons: one, often definitions differ: if you look for the "largest cave in Canada" do you mean largest by length, by volume, by height, or by which other criteria? If you want to know the height of a mountain on the moon do you mean the relative height compared with the deepest point "near" it, or do you mean the height above a hypothetical sphere giving the average height of the moon (as we sort of do on Earth when we compare heights to sea-level); second, the discrepancies are often due to the fact that information comes from different times: it is very unfortunate that documents on the Web are rarely dated!

This, by the way, is the reason for the different judgement of the edibility of the mushroom mentioned above: it was eaten without known side effects for thousands of years; in 2002, suddenly two deaths seemed to be linked to the consumption of a dish made out of the mushroom. Whether the death of two persons makes a substance poisonous is very doubtful in itself: after all, we have people with peanut, fish, milk, etc. allergies (that can be deadly), yet we do not consider either peanut, nor fish, nor milk poisonous! Anyway, the isolated cases mentioned have caused newer entries on the mushroom to call it poisonous.

What can be learnt from this? (a) if various definitions are possible, the documents should make this clear: this is NOT a job for search engines but for authors of documents; (b) all documents should be clearly dated; (c) the date should be considered as part of the ranking algorithm in search engines. Note that if I search for a meeting, an algorithm, a tool, etc., I am likely to be more interested in more recent ones than in ones 10 years or further back!

There is even a darker side to it. We often warn (young) people today that they should not put up too much personal information on social sites like Facebook, because it can be used against them sometimes in surprising ways. But this is not a problem of Facebook alone: if you have ever left a picture of yourself with a girl/boy friend on a photo site, that site may not allow you to ever remove it again. Yet that picture may prove an embarrassment when you have married someone else at a later stage. Although some search engines do allow to ignore entries that are a number of years old (Google has such a parameter), not all search engines allow the deletion of old entries and those that do, do not advertise it much, so that the average user is not aware of them.

There is one other issue concerning search engines: web search engines traditionally work with a group of input words, connected by "or", "and" or "not". A more linguistic approach (natural language queries) was already taken in [4]. Natural language queries have been allowed in this electronic dictionary now for over 5 years (which in its full form remains fairly expensive, unfortunately). One of the

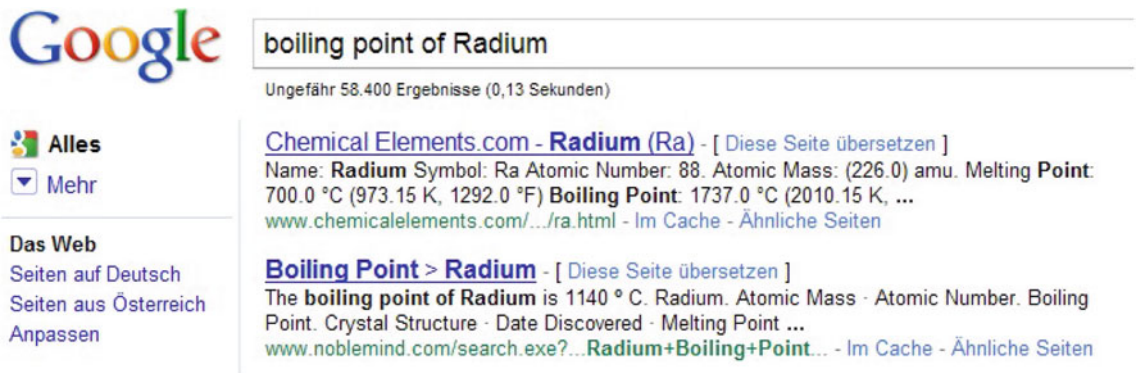

Fig. 1 Boiling point of radium

first tricks was to observe the word at the beginning of the query: "Who" is clearly asking for a person, "Where" for a location, "Why" for an explanation", etc. Consider as example the query "Who was York": this immediately excludes all cities with name York (Wikipedia lists some 50!), and other areas (music, companies, etc.) where York is used. There are still some 60 reasonably well-known persons named York, yet the resulting list is now bearable, and the person most often quoted, the explorer on the Lewis and Clark expedition can be ranked first in a good system. By delving more deeply into language understanding the linguistic group in Saarbruecken, Germany, has improved linguistically based searching considerably and has led to the establishment of a new large research cluster in this area called "Multimodal Computing and Interaction" see [24]

Despite this fact most search engines are still based on words, albeit more and more cleverly. Inputting "Who was the inventor of the toothbrush?" (in Google) turned into "Who invented the toothbrush" (i.e., some linguistic analysis is employed). The Google result gives what it seems is a reasonable answer ("No exact date known. . ." but then continues to give lots of additional information, like that the first mass production was started by a William Addis in 1770). That linguistic analysis is taking place rather than just using the important words is evident: when inputting "toothbrush inventor" into Google we get somewhat different answers.

The search engine Bing with input "Who was the inventor of the toothbrush?" finds William Addis in 1770, but gives very different answer on "Who invented the toothbrush". This seems to indicate that less language analysis is applied in Bing! The results above show that there are (mushroom case!) not just discrepancies within a search engine, but results also differ a lot depending on how a question is formulated. Further, discrepancies between different search engines can be quite serious! Often there is no easy way to determine who is right. In essence, one can trust the result of a query only if one can trust the source.

The question "Who was the physicist born in Vienna and died in Italy?" does not work well with Google. The reason is clear if one looks at the search results: the search is text based, so Google finds all Vienna physicists. Since Schrödinger worked (but did not die) in Italy at some stage his name pops up quite early, i.e. the verb "die" is ignored. Bing actually finds Boltzmann better than does Google, and provides interesting further information, yet its search is also clearly word-based. In the system which we will mention in the last Section of this paper, since documents have meta-data associated with them, Boltzmann is found immediately. Since general search engines cannot assume the presence of systematic meta-data, they either have to work with words or have to dig deeper into natural language understanding! But even if they do, how can we trust the result (see toothbrush example).

Summarizing this section: It is apparent that the major search engines do not employ deep language-analysis tools yet, are generally not good in allowing to narrow down large query sets, do not seriously try to reduce redundancy and do not take dates ("time-stamps") sufficiently into account: hence, much remains to be done to satisfy users. Due to the importance of search engines further progress can be expected, however. From a user point of view it is important that with Bing, with the completely different wolframalpha, with [3] and many others attempts new ways of searching; hence even Google will not be able to rest on its laurels. The authors wonder when the first search engine will become public that only searches sites with semantic data and a guaranteed level of reliability: it could turn the Web from a valuable but doubtful resource into something of much greater value than is offered to us today.

3 Special purpose encyclopaedias and dictionaries

There are thousands of free encyclopaedias and dictionaries on the Web. Some give only limited access free of charge but ask for payment for "premium use" or such. Some (typically medical encyclopaedias) are only available for closed user groups (certified physicians). One of the first such medical Web encyclopaedias was [12], offered by Bohmann Company Vienna for a number of years free of charge. However, like even much larger encyclopaedias (e.g. in Germany Brockhaus and Meyer, the latter available online free of charge for many years) most universal electronic encyclopaedias have disappeared or are only offering limited information for free, due to the pressure of free information, particularly from Wikipedia. For a while it seemed that Britannica would also give up completely, yet the current [7] electronic premium version is quite remarkable although only parts are free.

However, [18] shows that both general purpose ("universal") free encyclopaedias still do exist and that there is also a very large number of specialized encyclopaedias and dictionaries.

Clearly, Wikipedia has eroded the commercial base of general purpose high quality encyclopaedias to some extent, at least for the time being. While this has been deplored by some critics like [15], claiming that this is the beginning of a road to mediocre material, a vast number of persons believe that Wikipedia is such a valuable and also high-quality tool that the demise of commercial products is quite acceptable.

Although the authors of this paper have some points of criticism concerning Wikipedia they also are critical of traditional encyclopaedias for a reason that is often overlooked: the typical encyclopaedia of the twentieth century was an alphabetic arrangement of topics in an "objective" way, thus

reporting the "truth" about an event, a person, an idea, whatever.

We believe such a concept is basically flawed. Everyone agrees that if we look at a material object (as sculpture, a mountain, a house, any object you can think of) we can get a proper impression of the object only by seeing it from different views. This does also apply to non-material objects such as ideas, or personalities, etc., yet in general this is less explicitly acknowledged. But if we can only understand a complex person, a complex idea or a deep concept by getting very much opposing views a single "compromise" or "wishy-washy" description of the issue will not be helpful. What is needed are a number of different reports on the same subject with pointedly different views.

Traditional encyclopaedias have tried to live with this by having pro and contra views, yet there was always an author or team of authors behind each entry with a certain point of view, colouring the presentation. It is our belief that in future collections of encyclopaedic type this has to be avoided.

Similarly, it has to be avoided that encyclopaedias present an issue from a single point in time, since this often hides important issues. We have argued earlier that information consolidation in the sense of putting thousands of queries into a number of clusters, each cluster representing a certain point of view is important. Applying this to encyclopaedias we are very much in favour of presenting not one entry for a particular topic, but a number of entries. Information consolidation (more recently the terminology "information integration" seems to be winning) in the sense that all information about one topic has to be somehow combined into one entry gives rise to compromises, to what we called "wishy-washy" contributions earlier. Quoting [5]: "Information consolidation is bad for democracy because democracy requires an educated populace, one that has the ability to understand different points of view and think critically about complex issues, in order to flourish."

Let us explain this with one simple example. In the eighties of the past century Europeans were so worried about the extinction of interesting varieties of tropical wood that the import of certain types and objects made thereof was forbidden. A typical European encyclopaedia of 1985 would report this fact with some pride, showing the concern of Europe for maintaining variety in nature. However, since the import of tropical wood was not possible any more that type of wood lost its inherent value. Because of this, large forests of threatened species of tropical wood were burnt down to make room for rice fields that would yield at least a bit for the local population. Thus, the well-intended effort to protect tropical wood produced exactly the opposite of the desired effect. And a European encyclopaedia of 2002 reported (a) that certain types of tropical wood are endangered and (b) that local population was continuing to destroy it. The reason for this was (often) not mentioned. To avoid such misinformation by not providing enough information seems to be important. Sites like Sqidoo, HubPages or Helium.com do allow to present rather different views on the same topic, different from what is done in Wikipedia.

This leads us to a critical analysis of Wikipedia. It turns out that Wikipedia might well be a step in the right direction but that some changes would indeed increase its value still further.

4 Wikipedia

Wikipedia is certainly one of the big successes of the "Wisdom of the Crowd" paradigm as described in [27]. According to [30] some 400 million persons are using Wikipedia nowadays.

Over time, many weaknesses have been pointed out: in addition to inadvertent errors there have been cases of deliberate spreading of false information including defamation of persons, blown out of proportion description by paid or unpaid fans of some notion or person, hidden advertisements, or discrepancies in numbers reported: In some report on some country the population of city A would be mentioned at a number x, while the report dedicated to city A would mention a number y, potentially because census data from different time periods had been used. Another troublesome aspect is that the same event might occupy much space in some language version of Wikipedia, but may be quite short in other languages versions. Worse, the inventor of some device D might be person A in one country, and Person B in another country.

However, having said all this it is also clear that the average quality of contributions is quite good that the control of many readers is working to a high degree. It also must be understood that editing, censorship and correction procedures can be quite different between various language versions of Wikipedia and that rules are not carved in stone, but keep being improved. Here is an example: in the English Wikipedia it was initially possible to write completely anonymously. After a famous slander case this was given up. We quote from Wikipedia itself:

> "The ***Seigenthaler incident*** was a series of events that began in May 2005 with the anonymous posting of a hoax article in the online encyclopedia Wikipedia about John Seigenthaler, a well-known American journalist. The post fabricated statements that Seigenthaler had been a suspect in the assassinations of U.S. President John F. Kennedy and Attorney General Robert F. Kennedy. The 78-year-old Seigenthaler, who had been a friend and aide to Robert Kennedy and a pallbearer at his funeral, characterized the Wikipedia entry about him as "Internet character assassination".

> "The hoax was not discovered and corrected for more than four months, after which Seigenthaler wrote about his experience in USA Today. The incident raised questions about the reliability of Wikipedia and other websites with user-generated content that lack the legal accountability of traditional newspapers and published materials.[3] After the incident, Wikipedia co-founder Jimmy Wales stated that the encyclopedia had barred unregistered users from creating new articles."

Thus, today, at least some versions of Wikipedia do not allow to write contributions unless some screening of the writer has taken place.

We have criticised that traditional encyclopaedias have only one entry for even the most complex topic, even if that topic cannot be presented by someone "claiming to have the truth" but only by presenting different points of view. Wikipedia is doing the same, yet it does allow to examine the thread of discussion that has led to the current result, thus giving much more insight than is the case in traditional presentations.

Much effort and research has gone into improving the quality of Wikipedia. First, the editorial process is quite complex and it is trying to assure high quality, see, e.g. [26]; sophisticated techniques to test the quality of a contribution by number of editors involved and a word-count (!) have been investigated, see [1], and there have been many attempts to compare information quality with other sources, a difficult undertaking since there is no agreed way to measure of information quality.

However, we feel that a number of crucial improvements are still missing to make Wikipedia to what it is now trying to be: the ultimate source of reliable information on any subject whatsoever. To prove our point we are in the process of establishing an undertaking where we try to reproduce what is good in Wikipedia, yet where the introduction of a number of additional features should help in achieving a new kind of quality. It would be unrealistic to do this on the scale of Wikipedia, so we have restricted the scale by only collecting information on a single small country and issues involving it, even omitting topical issues, but instead digging much deeper and using not only the community to generate material but also large existing bodies of information.

5 Our system

The system (technically a JSP WIKI with many plug-ins) has been officially in operation since October 2009. It covers only "Austriaca", i.e. items that involve Austria or Austrians in some way. At the time of publication the system that is accessible at http://www.austria-forum.org comprises about 260,000 "objects", an object defined as text-file, picture, panoramic-, audio- or video-file. Completion of the desired functionality and a first solid foundation information-wise is planned for end of 2013, at which point Austria-Forum will contain close to half a million objects.

It is important to understand the main differences between Wikipedia and the Austria-Forum:

(i) In the Austria-Forum the domain is restricted to Austriaca as described; it emphasizes information that has a high degree of stability. Thus, a biography of a former poet or the description of an event in history is well suited, a biography of a rising new star in politics is acceptable and results of the rescue of the Chilean miners in October 2010 (despite the fact that Austrians were involved in it in a critical way) or sports events of the last month have no place in Austria-Forum: "news type" information is left to the media. One reason is to avoid competition, the second is pragmatism: we do not have the resources to also cover all those items and the third is maintainability. Once all important historic facts about Austria, all mountains, flowers, animals, minerals, stamps, coins, etc., etc., are collected, maintenance is comparatively easy: the biography of a poet like Stifter needs no updates, nor does the description of building the first road over some alpine pass; and although flowers and animals might change a bit, the emphasis is a bit, i.e., keeping this up to date is a manageable effort, particularly since news reports by the Austrian Press Agency (APA) are analyzed (almost) automatically. This in itself is an interesting undertaking: a full description of our techniques would be too long; hence a few words will have to be sufficient: APA sends us only contributions in certain categories (persons, institutions, buildings,. . .). Each contribution comes with a title and a few keywords. The title is compared with titles already existing (if not, special heuristics are used) and is considered "irrelevant" or "likely to be interesting" based on title, keywords and some heuristics. Two examples will give at least some impression, we hope. Suppose the title is "Fischer" (Austria's president) and the keywords are "Travel" and "Japan". This will be an entry to be discarded (it is likely to just report on a run-of-the-mill state visit.) Suppose, however, the title is "Opera-House" and the keywords are "Fire" and "Destruction" the entry will be considered relevant and will be added to the history of the Opera-House at issue. Note that we do not have a 100 % rate of success with our methods, but a very high percentage: we get this, mind you, with a certain amount of human intervention. We believe that the collections of heuristics we apply with some human control can be further refined by machine learning technologies but we are at an early stage in using machine learning (also in other aspects of the Austria-Forum) and hence have to leave this for a future report.

In passing let us note that objects in Austria-Forum have one important meta-data field called "control". It can be set to "never" (if the item never needs to be controlled") or to

"sometimes" or "often." Most entries have the value "never", some (like cities) have the value set to "sometimes" (so the system sends a reminder to check the entry for new buildings, current companies or such ever two years) or "often" (for living persons who still may change position, get a distinction, or die, etc.). Due to the fact (discussed later) that we often do not want to update certain entries (but have versions with various time stamps), maintenance of Austria-Forum once the basic inputs are there is managable by one person, and Graz University of Technology will provide this minimal support. Current negotiations with three foundations make it likely that a limited budget allowing to go beyond pure maintenance will be available.

(ii) The Austria-Forum distinguishes between approved main entries and general entries in the community section. In the latter, rules similar to Wikipedia apply, yet contributions can be upgraded and moved to the main entries section if the editorial board so decides. Main entries come from various sources.

Some have an author who has been screened (by an editorial committee) and whose CV is available to users, so that they have background information on who is writing what. Note that although screening of authors is done it is less critical than in the case of Wikipedia: first, authors are introduced (with CV and picture) and thus will not risk destroying their reputation; second, other qualified authors might write a report on the same topic but from a different point of view; third, any user can add comments, hence pointing out errors or providing links to other sources, etc.

This approach, to use only a select group of editors rather than the whole community has been tried before, in all cases known to us with little success: Citizendium was introduced by Lary Sanger (one of the two founders of Wikipedia) when he was dissatisfied with the quality of Wikipedia. Yet this system has been around for some 6 years and still has "only" 16.000 entries, mind you of good quality [6], but with very slow growth. The introduction of collections of "knols" (modules of information) by Google using paid experts [17], was hyped at the start by some as "Wikipedia killer", but is also not more than surviving in a new form 'annotum'. Other attempts like [25] that are sometimes also mentioned as samples of encyclopaedias with refereed contributions but are really more refereed journals than encyclopaedias and quite similar to electronic refereed journals like the 18 year old Informatics Journal of Universal Computer Science [13].

So why should we hope that the Austria-Forum could do better? The main reason is that the largest amount of information does not come from the community and from members of the editorial staff, but comes from archives, books and encyclopaedias. Of course in all such cases the source is known; hence the amount to which one can trust a contribution is known to the reader. The philosophy behind using existing books and archives is simple: why should we rely only on editors to describe a part of Austrian history, architecture, nature or what have you, when we have recognized authorities who have written books or collections of books on this matter? Why should we rely only on the community for pictures, when Austria-Forum has agreements allowing to use all pictures of the historic data base IMAGNO [11], pictures of the Austrian National Library and suitable picture from the ten million pictures (!) of APA [2]? As a matter of fact, while Wikipedia has to be mainly concerned with quality control of authors and contribution, the largest effort in Austria-Forum is dedicated to the intelligent addition of contribution and pictures! To be more specific, of the ten million pictures of APA Picture Desk only some 200,000 are of interest to the Austria-Forum (pictures on a Formula 1 race, an earthquake somewhere, a political rally, demonstrations in Greece, etc. are not). The selection of suitable pictures (that are not yet in Austria-Forum or are very similar to ones existing) is a formidable task, subject of much research and other publications, e.g. [16]. Current approaches are base on a mix of heuristics with moderate success, requiring continuous human intervention in many cases. Whether machine learning can lead to adjustments of the heuristics improving them is not clear yet. The list of partners of Austria-Forum is by now quite impressive [23].

The fact that the Austria-Forum includes encyclopaedias from 1836 (in preparation), 1884, 1914, 1952, 1966, 1994 and a current one, and other famous books, from the past to the present time, allows a kind of "time travel", a "mash-up time-wise", in addition to of course existing "geographic mash-ups".

(iii) It is also an aim to associate a date with each entry: not the last date of a minor update, but the date when the main entry was created. Note that this has two aspects: we hope to be able to, e.g. show pictures with sliders to view the change of a city, a glacier, a river, or other items, over time. We even hope to have a slider showing different points of view on various subjects. The "time-stamp paradigm" also means that if someone wants to do a major edit to an approved entry, this is not considered desirable. Rather, a new entry with the same name is created. Thus, ideally, you will not find an entry on "nuclear energy" but a sequence of entries like "History of nuclear energy in Austria", "Why nuclear energy is important", "Why nuclear energy is dangerous", etc: pointed and provocative contributions about nuclear energy from various points of view and written at different times. Ideally, you should not find a picture of our city Graz, or an essay about Graz, but photos of Graz at various times, and essays describing Graz at various times. (At the time of the last proof reading, in addition to basic contributiosn on Graz we do have now already 135 different essays on Graz written about different issues in time under "Damals in der Steiermark".) Thus, quite in contrast to Wikipedia, an essay on

Graz should no be updated, but retained as time capsule, and another time capsule added later.

Contentious views should not be merged in to one "wishy-washy" compromise, but should be available as separate contributions. Austria-Forum does not always offer pre-digested information, but rather information in the raw, yet consolidated in the sense that contributions on similar topics should be easily identifiable, well-grounded but with little redundancy.

In many instances Austria-Forum does not copy information, but just integrates existing information by links. However, information is always embedded in a fashion that the user-interface does not change. Most audio or video clips like, e.g. http://austria-lexikon.at/af/Wissenssammlungen/Video_Archiv/Vorarlberg_von_oben/Sommer/Bregenz do not reside on the Austria-Forum server, but are embedded as if they were. Exceptions are, e.g. all samples from the encyclopaedias of music that are indeed part of Austria-Forum, see, e.g. http://austria-lexikon.at/af/Wissenssammlungen/Musik_Kolleg/Beethoven/Eroica_Satz_1 and many of the short historical clips.

(iv) Since contributions have a source and a date, it is possible to quote them in scientific contributions, an open issue with Wikipedia contributions. Austria-Forum is interactive inasmuch as anyone can add comments to a contribution: many comments may lead some editor to even write a new version of the essay, leaving the old essay with all its idiosyncrasies intact. Other communication facilities are also provided to hopefully strengthen the spirit of community, and visualization tools are being developed to follow the doctrine of information consolidation/Integration explained in [31].

(v) We do not believe in providing a single encyclopaedia, but a substantial set of them covering various topics. The reason is that the search in Austria-Forum allows to not only be narrowed down to one area (a very desirable feature) but also to use available metadata. Note that Fig. 2 shows a form filled out with entries typical for a biography and indeed the search finds immediately the person at issue (Boltzmann). But the form (metadata) required to find a lake, a building, a flower, etc. would clearly have to look very different. Thus there is no single ontology/ meta-data structure that applies to all of the about 60 collections of knowledge, but rather different meta-data entries will be used in different collections. A powerful search mechanism allows cross-domain searches, but is only word- and keyword based, yet further cohesion is achieved by algorithms that (semi) automatically link related data.

(vi) We have added to the Austria-Forum a new kind of object akin to an e-Book. Those books are stored in a kind of bookshelf: the initial parts of a few rows of some historical books are shown in Fig. 3.

The books themselves behave more like real books than, e.g. PDF-files, yet they do offer advantages like searches (some even in old fonts and handwriting!) as one would expect from electronic substances.

We believe that information on the Austria-Forum will get its main advantage because of much information integration by linking relevant information in books and outside them together, via time and opinion boundaries, in a sense even allowing trips through time. Let us mention two examples:

Opening the fifth book on the second row of Fig. 3 and going to the village of Obdach shows Obdach around 1895 (Fig. 4).

The small elliptical icon on the left indicates that there is a panoramic view of today of the same village available. A click at that icon results in the picture shown in Fig. 5.

The panoramic view does allow, as one would expect, to zoom in or out, to pan, and to tilt the picture.

Thus, Austria-forum allows to jump backward and forward in time, a feature that will be dramatically expanded over the next 2 years.

One other typical example would be if you type "Strudengau" into the search field of the Austria-Forum and select the "Heimatlexikon" entry: this gives you a description of the segment of Danube above Vienna which once was consid-

Fig. 2 Searching in Austria-Forum using meta-data

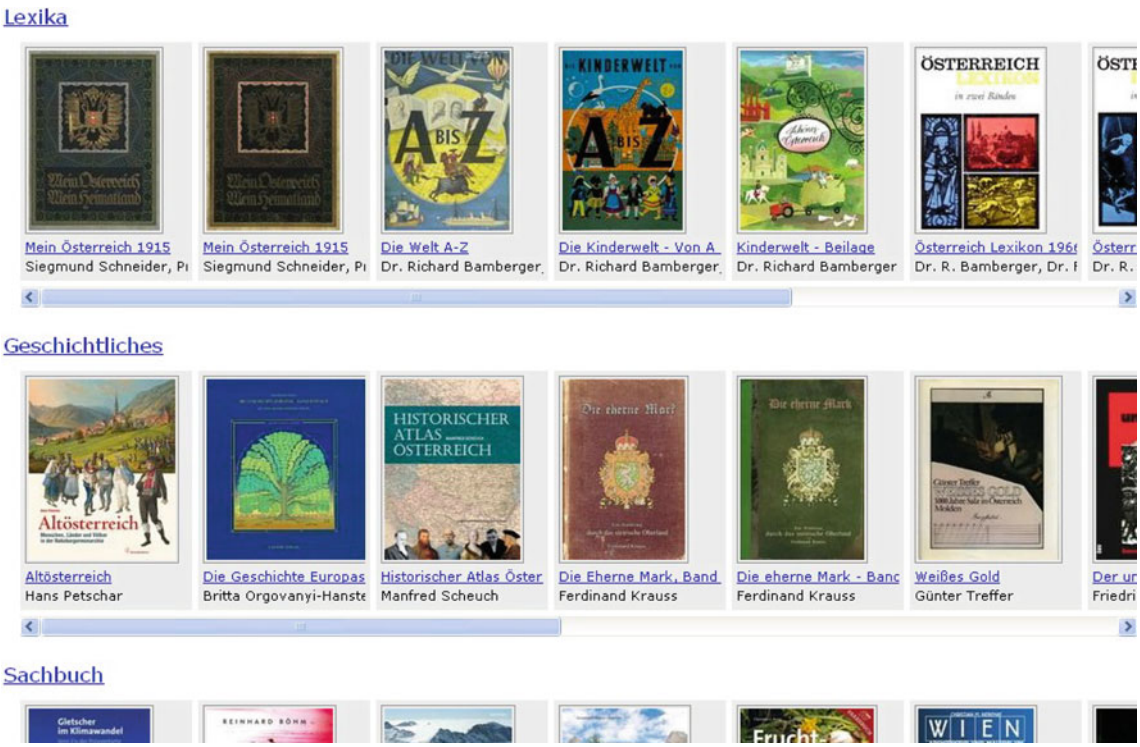

Fig. 3 Part of one of the book-shelves

in einem adeligen Hause, die Falb bald hierauf antrat, verschaffte ihm eine unabhängige Existenz. Falb gründete in dieser Zeit die populär-astronomische Zeitschrift „Sirius", welche er bis 1877 dirigierte. Bald darauf publicierte er das Werk: „Grundzüge einer Theorie der Erdbeben und Vulcanausbrüche", worin er zum erstenmale seine Theorie über den Einfluss der Mondnähe auf die Erdbeben-Erscheinungen aufstellt. In den

Obdach.

nächsten Jahren finden wir Falb an den Sternwarten und Hochschulen zu Wien und Prag, woselbst er die Vorlesungen Hochstetters, Hausteins, Dureges und Mays über Geologie, Physik, Mathematik, Planetenberechnung etc. frequentierte. F. trat hierauf zum Protestantismus über. Die Erdbeben von Belluno 1873 und der Ausbruch des Ätna 1874, welche Ereignisse Falb mit großer Bestimmtheit vorausgesagt hatte, verbreiteten rasch den Ruf Falbs und bestimmten denselben, seine Theorie wie folgt zu präcisieren:

Fig. 4 Obdach some 115 years ago

ered dangerous because of rocks in the river all over. It shows you a very old picture, some pictures dating back 100 years and a link pointing to a painting done in 1820. Clicking at it, you end up on the correct page of a book with that painting that you now can manipulate. But more important, a further link to an old encyclopaedia is shown and clicking at it you arrive at the correct page of that encyclopaedia. All together it gives you a good impression of this stretch of Danube across a few hundred years.

Allowing users to add their personal (or public) remarks and links will turn this new kind of object [22] into a valuable tool and into a solid basis for well-grounded discussions. Most important, it will have much more the feeling of being a genuine digital library than most previous effort, where the "emotional book feeling" has been largely ignored.

For readers eager to try this out look at, e.g. http://www.austria-lexikon.at/ebook/bookshelf/ and click at the first book on the shelf!

The interactive books are loaded double page by double page, i.e. internet connectivity is required all the time. A complete download or printout is only available as option. Usually, only one double page is visible at a time, and the direct print function is disabled. This is why even best-selling books are offered free on the Austria-Forum: the idea is that books will not be read this way, but just casually perused, maybe in this manner increasing the sales of the book in printed or in an electronic form that can be perused without internet connectivity.

(vii) In summary, Austria-Forum is not one encyclopaedia but a collection of such, based on some contributions specifically written for it, but mainly based on high-quality

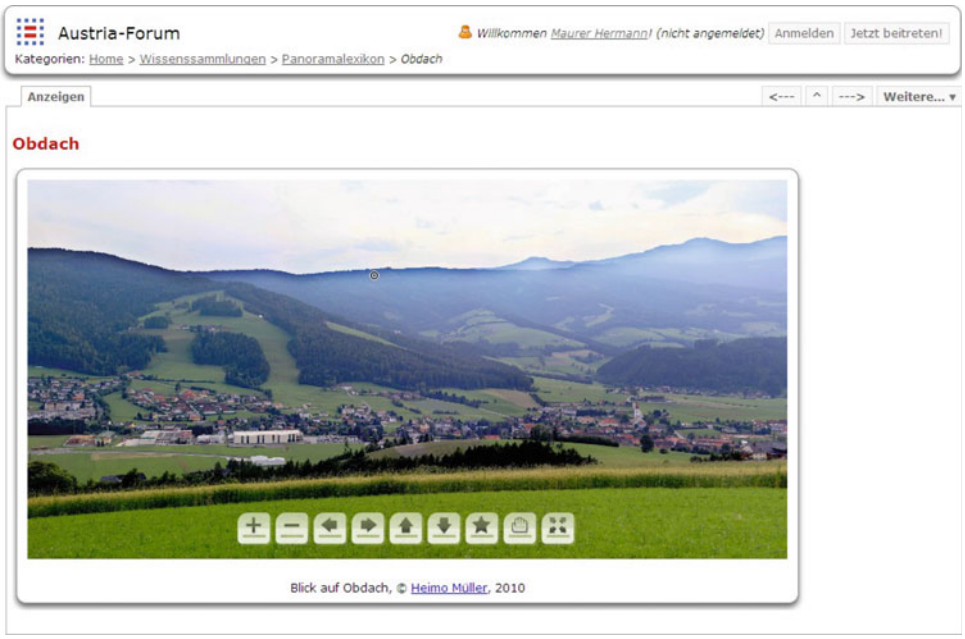

Fig. 5 A panoramic view of the village of Obdach

material already available in the form of archives, books, etc. This allows to dig quite deep in some instances, deeper than in any general purpose electronic encyclopaedia, even the size of Wikipedia. Thus, Austria-Forum does not compete with the German Wikipedia (Austria-Forum is mainly but not exclusively in German; particularly languages of the pre-1918 Austria will play an increasing role), but is rather an addition. It does not compete with the huge European Library project, since this is a big portal [10], tying libraries and museums together, but each library and archive (unfortunately) with its own interface. Also, the fact that it is a portal gives access much beyond what Austria-Forum or Wikipedia can do, yet redundancy is very high. Information on, e.g. some painter will be found in dozens if not hundreds of libraries. In this sense the European Library Project is a very valuable tool for researchers but not so well suited for a general audience.

6 Experiences and preliminary evaluation

Austria-Forum is much based on the premise to provide information from large existing resources, albeit encouraging both specialists and the community to contribute, criticise and add information and comments.

To achieve substantial quantity (for the moment concerning "Austriaca" only) it has been necessary to develop new methods to import data from various sources so that not too much redundancy will occur. The idea to import books as such (yet full-text searchable, with group annotations and the like) without exorbitant effort has been a challenge. However, by now uploading typical books (including the scanning and OCR involved if not a available as PDF file) has been reduced to a matter of hours and hence has turned out to be THE way to rapidly build up a data-base of high-quality material. It is the intelligent cross-linking of books that is still in its infancies, where we invest much research and where the community might be most helpful.

Note that many archives and libraries all over the world are in the process of digitizing parts of their holdings. A sizeable number of institutions in Austria is happy that their material is hosted for them in a coherent fashion on Austria-Forum.

It has been of critical importance to gain the backing of the Federal Ministry of Education of Austria to make sure that Austria-Forum will become one of the main and quotable sources in schools and universities.

As part of our research work we are constantly carrying out usability studies with the server and will report on some big surprises emerging from such studies separately. Just to whet the appetite: visible links are considered dangerous to fluent reading and understanding; hence we are experimenting with showing that a link exists only by mouse-over, and we have already integrated a search function that is activated by the double click at any word.

The growth of size form 90,000 objects at the opening (October 2009), to 150,000 (by October 2010) and 260,000 (by May 2012) suggest a growth rate of over 50 % annually. By increasing the efficiency with which we are dealing with external archives and books we believe this growth rate can be maintained for at least two further years, amounting to some 500,000 objects by end 2013. With some 1.2 million different users within the last year and a slow but

steady growth, it is likely that Austria-Forum will be one of the major sources of information on Austria. However, as Jimmy Wales said on May 27, 2010: "More Quality, less quantity", see http://www.businessinsider.com/jimmy-wales-wikipedia-future-2010-5. Hence the eventual test of success will be the amount of material that is quoted from Austria-Forum, e.g. by the number of ingoing links. That the number of links from the German Wikipedia is growing rapidly gives some reason for optimism.

7 Conclusion

It is an accepted fact that we are going to use material on the web more and more. In this paper we have analyzed a number of ways how to make reliable information accessible. We have argued that no approach is without its flaws. We have further explained a new approach that is currently developed that we hope will be a major contribution to handling the flood of information in what will look more and more like a physical library, yet comes with all functionality expected from a large electronic corpus.

References

1. Andera, M., Stein, B., Lipka, N.: Towards Automatic Quality Assurance in Wikipedia. WWW 2011 Poster (2011)
2. APA Picture Desk: http://www.picturedesk.com/bild-disp/apa/de/home.html (2011)
3. Blekko: http://blekko.com. Accessed 23 Nov 2010
4. Brockhaus.: Der elektronische Brockhaus. Mannheim, Germany (2006)
5. Cangie, R.: http://robinoula.com/media-and-technology/information-consolidation-why-aol-huffpo-is-bad-for-democracy/ (2008)
6. Citizendium: http://en.citizendium.org/ (2011)
7. Encyclopaedia Britannica: Electronic Version. http://www.britannica.com/. Accessed 30 Nov 2010
8. Encyclopedias: http://www.encyclopedia.com/. Accessed 25 Nov 2010
9. Eissen, S., Stein, B.: Intrinsic plagiarism detection. In: Advances in Information Retrieval. Lecture Notes in Computer Science, vol. 3936, pp. 565–569. Springer, Berlin (2006)
10. Europeana: http://www.europeana.eu/portal/ (2011)
11. IMAGNO: www.imagno.com (2011)
12. InfoMed Austria: Medizinische Informationen. Bohmann, Vienna (1999)
13. JUCS: http://www.jucs.org (2011)
14. Kappe, F., Maurer, H., Zaka, B.: Plagiarism: a survey. J. Universal Comput. Sci. **12**, 8 http://www.jucs.org (2006)
15. Keen, A.: The cult of the amateur. Double Day (2007)
16. Korica-Pehserl, P.: Semi-automatic information retrieval and consolidation of pictures from large databases. JUCS **17** (2011, to appear)
17. Knol: http://knol.google.com/k (2011)
18. List of Encyclopaedias: http://en.wikipedia.org/wiki/Listof_onlineencyclopedias. Accessed 30 Nov 2010
19. Maurer, H., Kulathuramaiyer, N.: Fighting plagiarism and IPR violation: why is it so important? Inf. Services Use **27**(4), 185–191 (2007)
20. Meyer zu Eissen, S., Stein, B., Kulig, M.: Plagiarism detection without reference collections. In: Advances in Data Analysis. Studies in Classification, Data Analysis, and Knowledge Organization, pp. 359–366. Springer, Berlin (2007)
21. Mihalcea, R., Corley, C., Strapparava, C.: Corpus-Based and Knowledge-Base Measures of Text Similarity. American Association for Artificial Intelligence (2006)
22. Mueller, H., Maurer, H.: A new approach for e-books for teaching and learning. In: Proc. ED.MEDIA 2011, Lisbon, July 2011 (2010)
23. Partners: http://www.austria-lexikon.at/af/Infos_zum_AF/Partner (2011)
24. Saarbruecken: http://www.mmci.uni-saarland.de/en/start (2011)
25. Scholarpedia: http://en.wikipedia.org/wiki/Scholarpedia (2011)
26. Stvilia, B., Twidale, M.B., Smith, L.C., Gasser, L.: Information quality work organization in wikipedia. J. Am. Soc. Inf. Sci. Technol. **59**(6), 983–1001 (2008). Online version at http://onlinelibrary.wiley.com/journal/10.1002/(ISSN)1532-2890
27. Surowiecki, J.: 2005. The Wisdom of the Crowds. Achor Books (2006)
28. The Austrian Ecnyclopaedia: AEIOU. http://www.aeiou.at. Accessed 30 Nov 2010 (1995)
29. Weinmann, J.J., Learned-Miller, E., Hanson, A.R.: Scene text recognition using similarity and a lexicon with sparse belief propagation. IEEE Trans. Pattern Anal. Mach. Intell. **31**(10), 1733–1746 (2009). http://www.computer.org/portal/web/tpami (2009)
30. Wikipedia Foundation: http://wikimediafoundation.org/wiki/Home. Accessed 27 Nov (2010)
31. Wurzinger, G.: Data consolidation in large bodies of information. J. Universal Comput. Sci. **16**(21), 3314–3323 http://www.jucs.org (2010)
32. Zaka, B., Kulathuramayer, N., Maurer, H., Balke, W.-T.: Topic-centered aggregation of presentations for learning object repurposing. In: Proceedings of AACE E-Learn Vancouver 2008, pp. 3335–3342 (2008)