even attempts at actual language understanding). Recently, a major focus on the application of these sophisticated hybrid methods in large production textual databases by the DARPA TIPSTER [Harman, 1992] and TREC [Harman, 1993a; Harman, 1993b] projects has produced some impressive successes and may encourage their transition from research efforts to more broadly deployed systems. The difficulty with all of these sophisticated methods, however, is that their operation is incomprehensible to almost all users. It is very difficult to predict what they will retrieve and what they will ignore. Some critics of these approaches have termed them "information retrieval as magic". These technologies raise very real integrity and access issues in that they work reasonably well often enough to be useful but seldom work perfectly; worse, they fail drastically in a reasonable number of cases. And information seekers not only have no idea what these retrieval systems are doing, but very little sense of when they are or are not working right; and, as they move from one system to another (as will be increasingly common in a networked information environment) they also have no sense of the specific features and idiosyncrasies of a given retrieval system. And, unfortunately, little effort seems to have been invested in researching effective means for these systems to explain and document their processes to their users; such features would help a great deal.

To some extent, these sophisticated "voodoo" retrieval systems have been kept from the general public by groups like librarians who are sufficiently infomation-retrieval literate to recognize the problems and be alarmed by them. The general public won't care; as soon as these developing technologies become effective enough to provide a useful answer most of the time, the public will accept them (and swear at the "stupid computers" in cases where they don't work), unless we see an unprecedented rise in public literacy about information and information retrieval techniques. The unreliability of probabilistic and statistically based retrieval algorithms is today not a problem that the public understands; without such understanding they may well become victim to their limitations simply because they are easier to use than more traditional, deterministic approaches.

6. Access to and Integrity of the Historical and Scholarly Record

One can consider a printed work as knowledge bound at a given time. For example, an encyclopedia published on a certain date represents the common wisdom of society about a number of topics as of some point in time. Indeed, old encyclopedias, obsolete textbooks, out of date subject heading classification guides and other literature represent primary databases for cultural research[37] and for understanding our culture's view of the world at a given time. The scholarly record in any given area, viewed as a series of frozen artifacts narrowly spaced in time can be viewed as such a historical record.

The same issue applies to mass media. The daily, weekly and monthly publications of popular journals provide a nearly continuous chronology of the shifting perceptions of any number of cultural issues. The selection criteria for what is published are themselves a very important part of the cultural record, and represent very definite

371 am indebted to Professor Michael Buckland for illuminating this point.

biases (in some cases, one selects information sources precisely for the benefit of those editorial biases). Further, as information technology has made publishers more **agile,** as we have moved more to broadcast media (where only the present exists, in a real sense, and it is very difficult to go back and look at the media's content at earlier points in time) and as means of monitoring audience response have become more precise and more timely, content can be changed almost continuously in response to audience interests and preferences rather than reflecting a consistent editorial position Indeed, this content shift may take place hour by hour in the popular media: one can envision services such as the Cable News Network (CNN) shifting perspective from one broadcast of the news to the next (every half hour) based on viewer feedback and sensitivities .[38]

In a real sense, electronic information resources invite an Orwellian, a historical view of the world. Consider an electronic encyclopedia that is updated weekly or monthly; entries for countries and political movements are freely replaced. Rather than a series of views of events fixed at specific times, the entire view of the world is now subject to revision every week or two. There is no a priori reason why the implementation of such an electronic encyclopedia must ignore the past, but this is the simplest implementation; overlay the obsolete with the present.

Within the database management system community a concept sometimes termed "time-travel databases" has been developing; these are databases that can be viewed based on their contents as of a given moment [Stonebraker & Kemmitz, 1991]. As the database is updated, older versions of database records are retained, along with information as to when updates were applied and when information is replaced. Records are not actually ever deleted in such databases; rather, an indication is stored that notes that a given record has become invalid as of a given point in time. Such versioning or time travel databases are still at the research stage, however, and most commercial DBMS software does not support the necessary range of functions to allow production implementations of databases that incorporate a historical record of database evolution. Further, even if software becomes available, there are substantial costs in disk storage and retrieval efficiency that must be paid in order to provide historical views of database content. Libraries, facing continued financial pressures, will be hard put to justify investment in these technologies. Yet the ability to retrieve the state of knowledge or belief about a topic at a given point of time is an essential element of the historical scholarly record, and indeed a critical part of the data needed for a wide range of research endeavors.

The shift from sale and copyright law to contract law also raises issues where integrity and access combine in complex ways. Once a library purchased or otherwise obtained a physical artifact (for example, through a donation) that it made part of its collection, this artifact became part of the library's permanent collection. With the replacement of the transfer of artifacts by licensing of electronic information, it becomes much more difficult for a library to maintain early editions, erroneous distributions and other

[38] Continuous news broadcast services such as CNN currently modify about 6-8 minutes of their coverage from one cycle of the news to the next, dropping stories, adding stories, or making editing changes to stories that are repeated from one hour to the next (with these editing changes not necessarily being necessitated by new news developments).

materials that may be a part of the historical record which the publisher of a given information resource is not necessarily eager to have generally available. One need not assign any malice to the publisher wishing to withdraw out-of-date versions of their publications from circulation; the publisher may be doing this with the best motives, such as ensuring quality control. A pharmaceutical database publisher may want to ensure that incorrect or obsolete information (which may, indeed, be dangerous—for example inaccurate dosage data) is corrected promptly and comprehensively. More generally, a the publisher of an electronic newsletter may simply want to ensure that the corpus of published material is as accurate as possible; there is an inherent conflict between quality of a published corpus and the accuracy of the electronic publication as a historical record. The integrity of the historical record becomes far more subject to the desires of the publisher.

Earlier in this paper, copyright was identified as a potential barrier to access in the electronic environment. In the context of integrity, however, it can serve a very valuable purpose for authors by providing a basis for the author to ensure the integrity of his or her words over time, and preventing later "amendments" or "corrections" to published works. The right to make changes, like other rights (such as republication or translation rights) is subject to negotiation between author and publisher.

Of course, by the same token, one can imagine situations where a publisher (for example, a government or some other entity) uses its license control over material to effectively rewrite the historical record; certain material is simply declared "inoperative" and removed from circulation.[39] This is another illustration of the extraordinarily strong position of publishers, authors and other rights holders in the electronic information environment and the loss of public policy control of the balance between the rights of creators (or rights holders) and the public.[40] Ultimately, there may well have to be a rethinking of the definitions and meaning of publication; in the print world there is a strong sense that once something is published, some copies are distributed and available to the public permanently. Even in cases where a lawsuit is successfully brought against a publisher for one reason or another, while a result of the judgment may be that the publisher ceases to sell the work and destroys existing stock, there is really no practical way to recall copies already sold. Publication in the print world is generally viewed as an irreversible act;[41] at least under some definitions of

39 While slightly outside of the main focus of this paper, one area that I find particularly interesting and troublesome in this context is control of the news. The primary record of historical events is copyrighted material owned by newspapers and the broadcast media. As the various trends discussed in this paper lead to a situation where less and less of this material is held by libraries, it raises the specter of situations where for whatever reasons the primary historical record of events in a given area might well become inaccessible to researchers.

40 Copyright is not the only issue here, however. For example, an executive order was signed during the Reagan administration that permitted the government to reclassify previously declassified material in cases where they were able to regain control of all copies of the declassified document. In a print environment this is quite difficult; in an electronic environment it would be much easier.

41 Because of this irreversible nature of the act of publication, the scientific community has had to develop the practice of withdrawing a previously published paper that is later been found to be erroneous, for example; this is accomplished by printing a notice in a subsequent issue of the journal. This is not necessarily very effective, since a reader of the withdrawn paper may be unaware of its status. In an electronic environment, it is interesting to speculate how a withdrawn paper would be handled. Would it continue to be distributed, but bearing a prominent notice that the author has subsequently withdrawn it, or

"publication" this is not the case in the electronic environment. The new electronic environment is likely to create a great demand for what are perceived to be neutral parties to maintain the historical and cultural record (and I believe that most people view libraries as falling into this category), as well as various forms of audit trails so that revisions of the "record" can be tracked and evaluated.

The ability of a library to acquire access to data rather than copies of information resources also threatens the historical record of scholarship and culture. One vision of electronic information resources calls for publishers to mount databases of journal articles on the network, rather than supplying copies of the information to libraries that subscribe the these journals. In such a world traditional journal subscriptions are replaced by contractual arrangements that allow libraries to retrieve articles from the publisher provided servers under various terms (pay per view, or unlimited access to articles from the journal through a "subscription" price, for example). This may represent an economy for libraries, in that they do not have to receive journal issues, and they do not have to pay for local storage space to house these journals. But what happens when a publisher decides that a given journal (or specific back issues of that journal, or even specific journal articles that appeared in the journal) are no longer being used enough to make it profitable to provide access to the journal on the network? Or, perhaps, the publisher goes bankrupt, or is acquired by another publisher, or becomes entangled in litigation? In such cases it is quite possible that the publisher will cease to provide access to some or all of the contents of journals without notice, and without any recourse by the library community; the back issues simply become unavailable, for example. In the old print world, if a library somewhere held these back issues, they would still be available to patrons through interlibrary loan, but in the new electronic environment, unless some library had made local copies of the material (thus losing out on the economies that make electronic distribution of the matetial attractive) this material could be lost to the user community for all time.

Copyright law may again provide a useful tool for ensuring preservation of the scholarly record. Historically, deposit of a copy of a work with the Library of Congress has been a requirement for establishing copyright protection; while current copyright law has removed this requirement, to a great extent, a return to such deposit requirements could help to ensure the long-term accessibility of electronic material.

In the print world archival access to material was the responsibility of the library and archival communities. In a world of licensed access to electronic information, libraries cannot unilaterally continue to accept and discharge this responsibility. It may well be that the networked information environment will call for a new compact of responsible behavior between publishers and the library community, in which publishers make a

---

would distribution cease? In the OCLC/AAAS ***Current* Clinical Trials** electronic journal, the reader of any article that has had subsequent corrections or comments receives a very prominent warning that such supplementary information exists; however, it is in some sense **easy** for ***Current Clinical Trials*** to make such linkages visible to the user, since the journal is not simply distributed as content but includes a integral OCLC-supplied viewing **interface. Unless the historical record is actually altered in an electronic journal (at least to the extent of indicating as a note in an article that a correction or withdrawal notice was later issued, and when) electronic journals distributed as pure content (without a user interface to make such** links) **are** likely to offer only the same weak ability to notify users of subsequent corrections that characterize the print publishing world.

copy of their material available to some organization serving (and governed by) the library community so that the library community can assure itself of continued availability of material. Or a publisher might agree that if it removes material from availability on the network, it will offer this material to some access provider of last resort that is financed and governed by the library community (perhaps a network analog of the Center for Research Libraries for example). But the problem here is that while it is reasonably straightfoward to find solutions in an environment of cooperation between libraries and publishers in which all parties behave responsibly, there is the constant threat of irresponsible behavior on the part of publishers, or of external, uncontrollable events giving rise to the loss of key parts of the scholarly record.[42] National attention to the role of national libraries or other organizations in ensuring the preservation of, and access to, the scholarly record, is of vital importance in gaining the confidence of the user community in abandoning printed formats for electronic ones.

It should also be noted that there is another, more crass, issue that is raised by the transition to a networked information environment in which publishers are the primary providers of their inventory. Print is an inherently distributed medium, whereas in the electronic environment a technically inept publisher might stand to lose their intellectual property holdings through various types of catastrophe like fire, earthquake, or corruption of a network server by computer hackers. While the user community would not loose the rights to their material, practical access to this material might well become permanently lost.[43] From a business point of view it might mean that the publisher went bankrupt, but from the broader perspective of the scholarly community, it means that the material is lost and is no longer a part of the scholarly record. Given the numerous relatively small publishers, such as professional societies that issue one or two journals, loss of information due to failures of the publisher to adequately back up or protect their material should be viewed as a very real issue.

Natural disasters and business failures are not the only issues. As libraries move from providing access to their own local physical collections to a set of networked resources international issues must also be considered, for example. One can readily imagine situations where a national library in a foreign country provided access to the majority of the literature related to that nation, until suddenly some international political problem (an obscenity dispute, a war, a change of government, or whatever) caused that national library to cease providing the information in question. Even if there was no central point of control such as a national library access to information provided in one nation could be cut off for other nations by government action. Access may not just be interrupted; more subtle changes are possible. Imagine a fundamentalist religious government taking power in some nation; they might order the destruction of some

---

42 The issue of the scholarly record in electronic form should not be viewed in an entirely negative light. Today, in print, retractions and corrections probably rarely reach those who read the original article. In an electronic environment where we can track who has read (or downloaded) a given paper, the possibilities for disseminating retractions or corrections to the readers who most need to be aware of them is greatly improved.

43 Some publishers have argued that downloading in the Internet environment implies that there is always likely to be some copy of a publication stored on some individual's workstation, and that in this sense electronic publication on the Internet is also an irrevocable act. But, I would suggest that there is a great difference between continued access to material by some random member of the scholarly community and continued access by an institutional agency (such as a library).