the courts have largely upheld community consensus about what is reasonable, and sided with the consumer; any other choice would be impractical, to say the least. We will face many of the same issues in digital multimedia.

5. The Role of Secondary Information Sources and Automatic Indexing in Access to Electronic Information

Today, universities and other organizations are licensing and mounting abstracting and indexing (A&1) databases as adjuncts to their online catalogs; these databases provide library patrons with logical access to the journal literature (and sometimes also book chapters, books, technical reports, and other material) in a given discipline. A&1 databases contain citation information—authors, titles, journals of publication, page numbers—for material such as journal articles; in addition, they often include subject headings or other access terms assigned by indexers, and sometimes also abstracts of the contents of the articles, book chapters or other materials. Abstracting and indexing database records fill a role similar to that of cataloging records for books in a library catalog, but often provide more information about the work than a library catalog will. In general, library catalogs focus on monographic material-books, maps, sound recordings, films-or entire periodicals (for example, recording the fact that the library had a subscription to a given journal); abstracting and indexing databases typically focus on articles in journals or chapters in books.

Currently one of the major challenges for libraries is bridging the gap between the intellectual access offered by abstracting and indexing databases and access to their physical journal collections (as described in their catalogs thorough records of which journals they hold); in future the abstracting and indexing database providers will also offer links to electronic publications directly. The compilers of these databases wield great power that is just now being fully recognized. The experience of libraries in mounting online catalog databases (which typically cover only the monographic literature held by a given library) has been that when only the online catalog database was available some patrons tended to use monographic material almost exclusively; other (arguably more sophisticated) patrons who recognized that the journal literature was vital to their discipline tended to reject the online catalog as irrelevant. Indeed, this reaction to online catalogs was one of the primary forces that motivated libraries to license abstracting and indexing databases to attempt to bring access to the journal literature into balance with the access that they already offered to the monographic literature. Now that A&I databases in various disciplines are readily available to library patrons[26] these effectively define the relevant literature in these disciplines both in their

26 A few Points should be made about the origins and development of abstracting and indexing databases, and the impact of their conversion to electronic formats. In the mid 1800s various individuals and organizations began to compile indexes to parts of the journal literature and market these to libraries; however, the size of the journal literature was sufficiently small until the early 20th century so that at least large research libraries could actually create article-level card catalog entries for articles in journals to which they subscribed. Thus, up until the early 20th century, the library catalog served as a record of material that the library held, and specialized indices served as a means of providing access to the entire published literature in an area (whether the library owning the index owned the material or not). With the explosion of publication during the later part of the 20th century, economic considerations forced libraries to abandon the cataloging of articles in their journals, and they began to rely exclusively on subject bibliographies of the journal literature to provide patrons with access to journal articles. Thus, the print analogs of abstracting and indexing databases are nothing new. However, these printed tools were

selection of journals to index and in their chronological span. For all intents and purposes, if material in a given journal (or even a given issue of a given journal) isn't covered in the abstracting and indexing database, it might as well not exist from the patron's perspective.

Thus, the processes through which the compilers of these A&1 databases select which journals to index, and which articles within these journals should be indexed, are effectively defining the literature in various disciplines. Most library users are unaware of the precise chronological or literature coverage of these databases, or the differences from one database to another (and note that a library generally selects only one database in a given discipline, typically based on a mixture of quality and cost considerations, due to the very high cost of licensing and mounting such a database); indeed most database providers are very vague about even stating their coverage and selection policies, which can be substantially complex. This confusion is compounded by the fact that these A&1 databases evolve over time, and revisit their selection of journals to index, and the indexing policies (i.e. cover to cover indexing, which creates a record for every item that appears in the journal, or selective indexing, which only

---

generally hard to use and were seldom consulted except by librarians and by scholars familiar with their organization.

In the 1960s the organizations that prepared these bibliographies of the journal literature began to employ computers to manage citation databases than were then formatted for print; as the cost of computers began to drop, they made the databases available for online access, either directly or through service bureaus like Dialog or BRS. The first such databases supported relatively well-funded disciplines like the biomedical and health sciences (for example, the MEDLINE database), general science (the Current Contents and Science Citation Index databases), engineering (the INSPEC database), or the business and financial communities (ABI Inform); access to these files was very expensive (sometimes hundreds of dollars per hour) and because of the high costs use of these databases was largely limited to researchers in commercial corporations or occasionally academics with grant support. Universities sometimes offered a very limited amount of subsidized searching (for example, a few searches per year for faculty, or a search or two for doctoral candidates working on their dissertations). Also, because the search systems were not only very costly but also very difficult to use, most searching was performed by trained intermediaries (typically librarians with special training). As a consequence, while these databases were important resources for researchers in commercial settings, they had an extremely limited impact within the academic community.

In the 1980s computing costs dropped to the point where universities could begin to license these databases at flat fees and mount them on local computers for unlimited use by their academic user communities, typically using software that was designed to support access by end users rather than by trained search intermediaries. Usage grew by orders of magnitude; for example, at the University of California, popular databases such as MEDLINE now support in excess of 100,000 searches *per* week by the UC academic community, and the availability of such databases began to have a major impact on university-based research and instructional programs.

The other point that should be emphasized is the very powerful impact of computer-based information retrieval tools in academic libraries. The experience with online **catalogs was that most users of the library found these automated information systems so much more convenient than the card catalogs they replaced that they would typically use the online catalog even if its coverage was less complete than the older card catalog because some material in the card catalog did not yet have machine-readable records that allowed this material to be represented in the online catalog. Similarly, while the printed abstracting and indexing tools were very difficult to use, the online versions of these tools (at least in conjunction with end-user oriented retrieval software typically** used when they are mounted at university libraries) make *the* electronic databases very easy to use, and these databases consequently gain very high user acceptance and quickly begin to serve as the primary-indeed often nearly the sole-means of access to the journal literature.

creates records for certain material in the journal, based on type of material or article content) for selected journals from year to year; just because an A&I database currently covers a given journal at a given level of detail does not mean that it provides historical coverage of that journal, or that it has always covered the journal at the same level of detail.[27] Yet users—at least those in disciplines which still take the published literature seriously, as opposed to disciplines that view the key literature as preprints, technical reports and other electronic publications—tend to regard the coverage of the A&I databases available to them as effectively defining relevant literature in a discipline.

In a very real sense, the challenge facing an author of a scholarly article under the "publish or perish" regime still commonplace in academia for print publication is to get *published;* whether anyone reads the publication is a secondary issue. In the evolving networked information environment, all evidence suggests that it is all too easy for anyone to share their thoughts with the networked community through self-publication. The challenge in the networked environment will not be to make one's writings available, but rather to get people to read them. This will assign an ever greater emphasis on the selection and coverage choices made by abstracting and indexing services, particularly those that are explicitly recognized by scholarly communities because (for example) they are provided by various scholarly societies.

On one hand it seems that this trend is encouraging. Greater importance will be assigned to reviewers and bibliographers of all types. A researcher in a given area may well be willing to pay for the bibliographies of important recent articles provided by major figures in his or her field. Reviewers for journals-currently normally largely

---

[27] Close examination of editorial policies for abstracting and indexing databases indicate that they are very complex and have considerable impact on what information the user locates and how they can locate it. Consider, as one example, a popular database that offers coverage of the parts of the computing literature. Basic records in this database include the author, title, date of publication, subject headings describing the contents of an article and related material. Some, but not all, database records also include abstracts. Some of the journals in the database are indexed "cover to cover", which means that descriptive records for all material *of certain types* appearing in the journals are included in the database—but this may only include news announcements and articles, and not letters to the editor, errata announcements for articles in previous issues, conference announcements and calls for papers, or other materials. Advertising is almost always omitted, even lengthy special advertising sections and the sort of quasi-editorial material like new product announcements that are often found in trade journals. For other types of journals only articles related to computing are included in the database; thus a paper in a journal like *Scientific American* would be included only if it dealt with computing. Since the database vendor incurs a significant additional cost for each abstract that is included in the database, abstracts are only prepared for some of the material, most commonly longer articles. The vendor also offers a supplementary extra cost product that provides full text for some of the material in some of the journals that are covered by database; journals are included primarily based on the ability of the database provider to negotiate an acceptable agreement with the journal publisher for the remarketing of the text of the material in electronic form. Within the journals that are supplied in full text form, the database provider again employs editorial policies to select only specific types of material for inclusion as fulltext, since for most journals the database provider must pay for scanning or rekey boarding of the material and thus again incurs substantial costs for material included. For some types of material there may be fulltext but no abstract. Now, add to this rather complex set of criteria for what is placed in the database the additional complexity that all of the editorial policies just described are subject to continual revision and fine-tuning.

The user of such a database is typically unaware of all of these subtleties. However, searching by subject terms will actually search a different, larger set of articles than those accessible when searching by full text or keywords in the abstracts, and the table of contents of a given journal issue as derived from this database are likely to be somewhat different than the contents of the printed journal.

unrecognized and uncompensated for their labors—may find their evaluations recorded in databases and assigned great importance. Those who edit, filter, and select may play a much more important role in the networked information world. But, at the same time, established arbiters of taste within a given discipline, such as the compilers of abstracting and indexing databases, may have a much greater role in describing the relevant literature of a discipline.[28]

A key question here will be the amount of diversity available. One perspective on the matter extrapolates from the existing compilers of abstracting and indexing databases: these are organizations that attempt to provide systematic and comprehensive coverage of the literature in a discipline. Developing these databases is a costly proposition; the creation of such a database is a major investment by a corporation or other institution. The other perspective uses the network to expand the reach of what has traditionally been interpersonal communication—someone passes an interesting article to a colleague. The individual-based filtering and selection services serve different purposes and in some ways are more valuable to information seekers increasingly pressed for time as they help such information seekers to locate key publications quickly. Here the model is more one of bibliographies and reader's guides, which can be produced for limited areas by a single specialist or a small cadre of experts with a fairly limited investment Of course, individual-based services are more subjective. One of the most attractive points about individually produced bibliographies and reader's guides is that it gives wider voice to major thinkers in a given scholarly discipline-the "geniuses", to use one reviewer's term, can reach beyond their immediate circle of students and colleagues to highlight what they believe to be particularly important works for the broader scholarly community. Both approaches will have their roles.

The entire issue of evaluation of literatures is controversial [White, 1989]. Some librarians and researchers (such as F. W. Lancaster) argue that this is one of the key contributions of librarians and of various reviewing services. Certainly, every library makes evaluations daily as part of its acquisitions decisions, but the often it avoids suggesting that one item it its collection is "better" than another once the evaluation decision leading to acquisition has been made. The argument has also been made that the standard review sources in many disciplines are at best very conservative: they only tend to cover material from certain mainstream publishers (and, indeed, in some cases they are owned by one of the major publishers in the field) and as such tend to reduce diversity and the introduction of innovative new material, in part because librarians at

28 Occasionally, one reads visions of future electronic libraries that include a very intensive reader commentary component. The idea is that readers will attach their reactions and comments to material placed in an electronic library by the primary authors. Effective realizations of such a framework have proved elusive in practice. There are too many readers, with greatly varying levels of expertise and objectivity. While broad-based reader commentary may be a useful thing to incorporate in future electronic libraries, I do not believe that it will replace the role of expert selectors and commentators. It is also worth noting that there are subtle intellectual property problems here. Will the general public be willing to contribute their comments on material for public access? Certainly, some experts will try to make income by providing such commentary; if the public at large emulates this, one has an administrative, legal and accounting nightmare. If the public does not, then one must ask why certain commentators are willing to share their thoughts on a work freely while other commentators are not. Some projects, such as Ted Nelson's XANADU {Nelson, 1988], have attempted to explore the compensation and intellectual property issues implied by a move from published works to a rich web of commentary that surrounds these works.

many institutions, overworked and/or lacking the necessary expertise to make an independent evaluation, will simply use the review sources as purchasing guides. It seems to be that the networked environment will increase diversity in reviewing sources, thought it is not clear to me that many librarians (as opposed to subject matter experts) will step up to the challenge of providing these new bibliographies, abstracting and indexing tools, and reader's guides.

The trend towards having large, costly abstracting and indexing databases define the "core" of a disciplinary literature is of particular concern in conjunction with visions of the future which place professional societies in charge of the canonical literature in a given discipline (see, for example American Physical Society document on the development of a future international physics electronic library [Loken, 1990]); the problem here is that while a given researcher who is out of step with the conventional wisdom in a given field may be able to make his or her thoughts available on the network, it is unlikely that anyone will find them. One can all too easily envision the "establishment" in a given discipline taking control of the definition of the literature in that discipline through the compilation of the de facto standard abstracting and indexing databases in that discipline. To a certain extent, the easy self-publishing that is possible in the networked information environment addresses these concerns, but as indicated earlier the challenge is not to be published but to be read, In cases when tenure and promotion are at issue, there is likely to be no near-term substitute for publication on a prestigious journal; but, when the objective is more communication with one's peers, the question is whether the developing tools for identification and discovery of networked information resources will provide an adequate "safety net" to allow self-published materials to be located and read by those peers.

Another aspect of the role of secondary information services is their role in author evaluation decisions—for example, tenure and promotion decisions for academic authors. Some of the more sophisticated universities are recognizing the potential for subjective bias that may be present in the traditional abstracting and indexing services, and prefer what are allegedly more quantitatively objective secondary services such as the various citation indexing offerings from the Institute for Scientific Information (ISI) such as Science Citation Index [Garfield, 1979]. Citation indices count the number of times that a given publication is cited in the published literature; it is only a short step from these to even more "objective" measures of quality based on the number of times that a given article is accessed (in electronic form, where this number of accesses can easily be computed); this raises fundamental questions about the privacy of searches and the uses to which searches can be put that are discussed in a later section.[29]

The growing power of abstracting and indexing services raises many questions that need to be explored, and places at least a moral responsibility on the abstracting and indexing services to exercise a very high degree of quality control (though the legal liability of such services, as far as I know, has yet to be defined; the general issue of legal liability of information providers is discussed in a later section of this paper).

[29] It should be noted here that citation rates are a somewhat controversial measure of the impact of publications. They are subject to "gaming" in various forms: repeated and extensive self-citation, or the development of tight circles of authors who continually cite each other's works [Pertiz, 1992]. Similar questions will undoubtedly apply to the use of measures based on the number of accesses to articles in the networked environment.

Consider the possible impact of an service that abstracts and indexes only selectively: for all practical purposes, by not including a given article, the service excludes that article from the literature of a discipline and makes it unlikely that researchers in that discipline will subsequently find the article in question. This at least is an editorial judgment, [30] consider the case where by some error an abstracting and indexing service "misses" an issue of a journal and all its contents (perhaps the issue was lost in shipment to the service, or lost by the service during its processing stream), or makes an indexing error which causes a publication to become unretrievable. Such omissions evidently do occur today in some of the major services that are used in contexts such as tenure and promotion decisions.[31] The entire issue of the quality of abstracting and indexing databases is quite complex and subtle; the interested reader might wish to examine the recent series of articles by Peter Jasco on aspects of this topic [Jacso, 1992a; Jacso, 1992b; Jacso, 1993a; Jacso, 1993b; Jacso, 1993c].

As we begin to transition from printed materials to electronic materials we tend to think of abstracting and indexing services (first in print formats, and now as electronic databases) as perhaps the primary means of identifying source materials. In fact, as more and more primary (e.g. full text, or source) material is available in electronic form, new methods of identifying relevant material will come into play based on various forms of automated indexing and full text searching,[32] [Salton, 1988]. This is inevitable for three reasons.

First, the human intellectual effort for abstracting and indexing is costly and the user community cannot afford or is unwilling to pay for people (particularly expensive people with subject expertise) to index everything, particularly in great depth. Even if an abstracting and indexing database is available which covers a given set of material, a library might offer that source material in electronic format but may have chosen not to license the abstracting and indexing database for any number of reasons (for example, because the library only holds a very small proportion of the material that is covered in

30 One of the reviewers of the initial draft of this paper raised a very interesting issues about editorial selectivity: based on the Feist decision, one might argue that a comprehensive, cover-to-cover indexing and abstracting service would have very limited protection under copyright, while a service that was more selective would find that their selectivity would justify stronger copyright protection. Copyright protections may well encourage greater selectivity.

31 In many disciplines there are multiple competing abstracting and indexing services. Publishers have less to fear than individual institutions from errors that are made by a single service; over the broad subscriber base, the multiplicity of services will make it probable that at least one service provides proper access to the publisher's materials. However, given the very high cost of acquiring an abstracting and indexing database, a given library or university will probably select a single supplier from the various alternatives available on the market; thus, for a given university community (within which, for example, a tenure decision is made) a single abstracting and indexing database will dominate.

32 Note that all of the issues raised already about the power of editorial decision making In the compilation of abstracting and indexing databases also apply to fulltext databases, whether created independently of A&I databases or constructed as extensions of these databases. Choices about what to include in fulltext will have to be made; some database producers may not choose to include full text of ail articles that they abstract and index, or more generally of all articles that appear in a given issue of a journal. And new opportunities for editorial bias appear: for example, a given service might exclude full text of articles that are critical of that service's performance or practices. Given that many users will be satisfied with the part of the published literature that is immediately available in full text electronic format, such editorial decisions can have a powerful impact.

the A&I database); in these cases there is no choice but to use information derived from the source material to provide access to it.

Second, while mechanisms based on full text may or may not offer "better" access to material [Blair & Maron, 1985; Tenopir & Ro, 1990], they certainly offer different access which is at least a useful complement to human intellectual indexing,[33] Access based on full text can be an excellent supplement to shallow abstracting and indexing databases (for example, those that provide little or no subject access). Full text access can help to identify documents that mention people, places or things that may not have been sufficiently central to the theme of the work to be recognized by an indexer or abstracter; in this sense they provide much greater depth of access. Further, there is a sense that full text access is less "biased" than abstracting and indexing services in the sense that human judgment does not come into play. Full text access can help users who are having difficulty with specialized controlled vocabularies typically used in subject classification in A&1 databases.[34] In situations where large textual documents are available online, the two techniques may be used together: first, search an abstracting and indexing database to identify relevant documents, then use full text based access techniques to identify relevant parts within these documents.

A third reason why full text based access is coming into wide use is because of the delay inherent in human intellectual indexing. Today, major (and expensive) abstracting and indexing services often run as much as four to six months behind the appearance of source material in print (and recall that the print material itself may be months or even years behind the distribution of preprints or manuscript versions of material with the "invisible college" community). As electronic dissemination of information increases the speed with which material is made accessible, these lags in abstracting and indexing will become increasingly unacceptable to some users of the material-particularly those interested in the most recent material rather than those performing retrospective literature searches. Full text indexing allows access through

---

33 The definition of the quality of a method of providing access to documents is a very complex and somewhat subjective area. In the information retrieval research community measures such as precision, relevance and recall are used+ssentially measuring how many of the relevant documents are retrieved by the access method, how many irrelevant documents are returned along with the relevant ones, and how many relevant documents are missed by the access method. Clearly, performing large scale comparative tests between different methods given these definitions is extremely difficult, since it requires that someone go through the entire database in order to determine the "correct" answer to the queries in order to evaluate the performance of the access methods being tested, because of the great variation in the kinds of queries issued by users (and the great variation in the performance of many access methods from one query to another), and because of the very subjective nature of relevant documents (since even experts do not always agree on whether a given document is relevant to a given query, and the judgment of the experts may still not agree with the judgment of a typical user who is not a subject expert). At the same time, we should recognize that while this is a hard problem, experimental results for various retrieval approaches on a wide range of large databases would be of enormous interest and value.

34 Full text access is also helpful in dealing with the fact that controlled vocabularies grow and change over time as new areas of interest emerge within a discipline and new discoveries and developments occur, but these changes tend lag substantially behind the events that cause them; this is a well-recognized problem with the Library of Congress Subject Heading controlled vocabulary list, for example. In most cases, the cost of updating subject terminology used in existing database records to reflect changes in the terminology is prohibitive; only a few very high quality databases such as the National Library of Medicine's MEDLINE do this. Often, one can find terms used in the abstract or full text of an article long before they become established in the indexing vocabulary of a discipline.

the apparatus of bibliographic organization to occur simultaneously with the act of (electronic) publication. It is also interesting to note in this connection that part of the problem is the size of the literature base that most comprehensive abstracting and indexing services attempt to cover (which goes hand in hand with their lack of evaluative information-they will help you find all the documents on a given subject, but not the three best surveys). If we see the development of large numbers of limited scope, highly selective and highly evaluative citation lists/bibliographies offered by subject experts as proposed elsewhere in this paper, we may find that these specialized lists are also much more timely than the traditional abstracting and indexing services.

Clearly some types of electronic material, such as newsfeeds, will require automated indexing; human indexing will introduce so much delay that much of the time value of the material would be lost. Multimedia information—images, video feeds and the like-present an additional set of issues. Today, we have very limited capabilities to perform useful computer based indexing of multimedia; general image classification is beyond current technical capabilities,[35] though automated transcription of speech (audio, or the audio track of a video segment) may become a production technology within the current decade, and this soundtrack could provide a very valuable access point for video information. Already, today, closed-captioning tracks in video material are being indexed and used to provide access points to broadcast information. And there is technology in experimental use that separates pictures from text in bitmapped images of printed pages [Lesk, 1991], or that attempts to detect scene transitions in video clips .

There are many different full text based retrieval methods [Tenopir & Ro, 1990]. The simplest provide searching for exact words that appear in the text, often with the option of including truncation (match only the beginning of a word), Boolean operators (i.e. AND and OR), and proximity operators (to require that two words appear close to each other) in queries. These full text access methods are easy to understand and predictable, although they often lead to rather poor retrieval results. Much more sophisticated methods have been developed in the information retrieval research community and are now starting to appear in large scale production systems. These range from statistically based methods pioneered by Gerald Salton and his colleagues over the last three decades[36] [Salton, 1988], which are based on frequency of word occurrence along with some very superficial language processing (word stemming) through much more complex techniques that combine statistical analysis with various syntactic and semantic analysis techniques from natural language processing (for example, analysis of parts of speech, identification of proper nouns or noun phrases, or

[35] Certainly there have been great advances in image recognition in very specific problem domains ranging from quality control in manufacturing processes through target identification for smart weapons systems, but more general problems, such as identifying the objects in a picture, remain intractable to the best of my knowledge. Further, really useful classification of images for general purpose retrieval involves a great deal of cultural knowledge as well as simply the ability to identify things: identifying a photograph of the President of the United States shaking hands with the Mayor of New York is far more useful than simply recognizing that a photo depicts two men shaking hands.

[36] In the past few years there have been a number of proposals for more sophisticated and computationally intensive statistically based indexing algorithms, such as the Latent Semantic Indexing techniques developed by Bell Labs [Deerwester, Dumais, Furnas, & Landauer, 1990; Foltz, 1990].