Subject classifications and thesauri have become more important than ever in the Web environment. Efforts made to organize information into subject classifications, or taxonomies, offer users the opportunity to substantially improve the effectiveness of their search and retrieval activities. This article continues earlier research on the development of a new definition of the field of information science and the creation of a "map" of the field showing subjects central to it and their relationships to those on the periphery. A case study describes the creation of a new classification structure (taxonomy) for the Information Science Abstracts (ISA) database, aiming to reflect and accommodate the rapid and continued technological and market changes affecting the information industry today and into the future. Based on a sample of some 3,000 ISA abstracts, two validation experiments were conducted by a three-member team comprising a database editor, a reference librarian, and an abstractor-indexer, who represent three of the major communities within the information science field. In the first experiment, the sample of abstracts was classified according to the proposed new taxonomy; after analysis of the data and revision of the taxonomy, it was revalidated and fine tuned in a second experiment. Indexer consistency measures obtained in this study were significantly higher than those found in previous studies.

The taxonomy resulting from this research employs the concepts, definition, and map of information science previously developed. It presents them in an organized hierarchical view of the field and thus makes a significant contribution to information science.

Introduction

Abstracting and indexing (A&I) services have long provided their users with aids to retrieving information from their databases. The initial products of many A&I services were printed publications organized in separate sections similar to chapters in a book and containing helpful author and subject indexes. Each section contained abstracts of articles on a broad subject. Users could therefore simply scan a section to find items of interest. The subject index portions of such A&I publications were generally based on controlled vocabulary terms developed at considerable cost and effort by information professionals with expertise in the discipline covered by the publication. Controlled vocabulary terms were frequently organized hierarchically into a thesaurus; the titles of the sections often became top-level terms ("main headings") in the thesaurus.

When A&I publications became available as searchable on-line databases, the controlled vocabularies and main headings were usually made searchable by the search service vendors. These added-value features provided valuable assistance to searchers, especially before the full text was available on-line. Many database producers expended considerable resources in training searchers (then mainly information professionals functioning in an intermediary mode) in how to use their thesauri. (One producer—the National Library of Medicine—initially required participation in a
Taxonomies in the Web Environment

One of the concerns a database producer must address is whether to expend the considerable resources required to produce a taxonomy and then to index the information using it. When commercial on-line databases first appeared, professional searchers clamored for the inclusion of thesauri and subject classifications to help them formulate their searches. Database producers responded favorably, and many of the early on-line databases included controlled vocabulary (descriptor) fields drawn from thesauri, as well as fields containing subject classification codes. The experiences of the past few years in the Internet world have made it clear that the effort to include such data is still worthwhile.

When databases of information (particularly in full text) first became available on the Internet, many users felt that thesauri and subject classifications were no longer needed and would go the way of horseless carriages. After all, the theory ran, if everything is available on-line in full text, one would only need to enter the appropriate terms into a search engine, and the desired information would be retrieved. Inexperienced searchers quickly discovered the fallacy of this approach to information retrieval when they were faced with result sets numbering in the millions of hits, with the desired information buried somewhere in them.

It was soon recognized that, far from fading away in the Web environment, subject classifications and thesauri have become more important than ever, and organizing information into subjects, or taxonomies, provided users with a significant improvement in retrieval. One of the earliest examples of the use of taxonomies was Yahoo!, which used trained information professionals to organize and categorize Web sites. Presently, in addition to a simple search box, most other search engines now provide an optional taxonomy that one can use for retrieval. The use of taxonomies has also spread to many large company intranets.

Initial Considerations

In a previous article, Hawkins (2001) described some of the history of Information Science Abstracts (ISA)1 and the development of a new definition of information science as well as a "map" of the field showing the subjects central to it and their relationships to those on the periphery. The work described here is an implementation and a practical application of the definition of information science and its relationship to related disciplines as outlined in Hawkins' article. We have used his portrayal of the discipline to develop a new taxonomy for information science—and for ISA in particular—which accurately reflects the field as it exists today. These concepts have also been actively used to select relevant articles for abstracting in ISA.

ISA's subject classification scheme (i.e., taxonomy) and controlled vocabulary last underwent modification in 1993, when minor revisions were made. Since then, as the information field experienced many major and traumatic changes, the taxonomy became outdated. It contained many outmoded terms, and it was no longer able to accommodate the rapid technological and market changes affecting the information industry. ISA's editors and indexers found it difficult to use; consequently, one can infer that users had even more difficulty. A major flaw in the previous taxonomy was that each top-level section contained a subsection entitled "General." Over time, many abstracts had been placed in those subsections because the field had advanced and the taxonomy had not kept up with the changes. The result was that the taxonomy became virtually useless as an information retrieval tool.

Figure 1, containing data taken from the master ISA production database, illustrates these observations. It shows the number of items posted in each section between 1992 and 2001. Note the wide variation in the number of items in the various sections. The section with the most postings (5.11, Searching and Retrieval) had over 2,700 items in it, and the one with the fewest (5.03, Supercomputers) had just four postings. Twelve sections had over 1,000 postings each, and 15 had fewer than 100 postings. These data provided strong evidence of the need for a new information science taxonomy, and the definition and map previously developed provided an excellent conceptual foundation for it.

Taxonomy Development Philosophy and Methodology

In these days of Internet search engines that are widely used by end users, the structure of a taxonomy must be clear and logical, and the taxonomy must be easy to use if there is to be any hope of its enjoying significant acceptance. It must also be dynamic, easily updated, and able to reflect rapid changes and technological advances. These principles are especially important for those users who scan the printed version of an A&I publication looking for items of interest.

If a taxonomy is well constructed, its terms can be used to advantage by on-line searchers who would like to make meaningful use of the subject headings as broad limiting terms in searches. Indeed, as we have already noted, many of the commercial search hosts construct an inverted index of the main headings for this purpose. [For example, the Dialog system places numerical designations of the headings in the SC (Section Code) or MC (Main Heading Code) field, with corresponding textual equivalents in the SH (Section Heading) or MH (Main Heading) field.] These search terms are extremely useful when one wishes to

---

1 ISA, one of the leading A&I databases covering the field of information science, is published by Information Today, Inc. It is available in print from the publisher, and on-line through several search services. For further information, see http://www.infotoday.com/isa/default.htm.
retrieve a large set of broadly related items and use it as a limiting criterion on other aspects of the search, in a "successive fractions" approach to searching (see Hawkins & Wagers (1983), for a discussion of this and other search techniques).

With these general guiding principles in mind, we collected the initial candidate terms for the taxonomy. The following sources were useful:

- Previous mappings of the information science field by Hawkins (2001) and others (see Hawkins’ article for references);
- ISA’s then-current list of descriptors;
- Lists of subject terms used by two other databases covering information science: R.R. Bowker’s Library and Information Science Abstracts (LISA) (now owned by Cambridge Scientific Abstracts), and H.W. Wilson’s Library Literature;
- The ASIS Thesaurus of Information Science and Librarianship compiled by Milstead (1998);
- Scope outlines for two of the central publications in information science, the Journal of the American Society for Information Science (JASIS) and the Annual Review of Information Science and Technology (ARIST); and
- Appropriate sections of thesauri used by several other non-information science databases that contain information science terms (e.g., the INSPEC database).

Candidate terms were grouped and organized into a preliminary taxonomy containing 13 main headings. Before being accepted, many of them were checked by searching for them in the ISA database and observing the number of postings retrieved as well as the context in which the terms were used.

Validation of the Taxonomy

Each of the authors of this article then independently indexed the December 1998 through May 1999 issues of ISA using the preliminary taxonomy. These six issues of ISA contain a total of 3,004 abstracts, so approximately 9,000 index assignments were made. Each abstract was assigned to the subject classification judged most relevant by the indexer. To simplify and speed this process, abstracts were given only a single classification number; in real life, of course, many abstracts have multiple classifications. Classification numbers were of the form x.y, where x is the “main heading” and y is the “subheading.” Abstracts for which the indexer felt that no appropriate classification existed were noted separately.

The classification assignments were assembled into a Microsoft Access database. If at least two of the indexers agreed on the classification of an abstract, then that assigned classification number was accepted as its correct indexing. Abstracts that were placed in different classifications by all three indexers as well as those that had been noted separately were reviewed, and the different points of view were either reconciled or appropriate modifications were made to the taxonomy to accommodate them.

Using the Access database, we were able to quickly obtain data on the distribution of abstracts in each of the sections (see Fig. 2). As a result, we were able to identify some extraneous categories and make the taxonomy more consistent by combining closely related categories into one. The distribution of section assignments, as shown in Figure 2, falls into three broad groups. Sections 1 and 8, which were concerned with basic information science research and electronic information systems, received over 35% of the assignments. A second group of sections (2 through 6), taken together, received about 45%, and the remaining 20% of the assignments fell into one of the other sections. This exercise was extremely helpful and revealing. The abstracts that could not be indexed pointed out several gaps in the preliminary taxonomy and the distribution shown in Figure 2 showed that further refinement of the sections would be appropriate.

Indexing Consistency

The validation exercise also allowed us to measure consistency among the three indexers, shown in Table 1.

Considering the full classification of each item, consisting of both main and subheadings, at least two of the indexers agreed on the classification of 70% of the 3,004 abstracts from these issues of ISA. Because the distinction between many of the subheadings may be rather small, we also looked at the consistencies when only the main headings are taken into consideration. On that basis, consistency rose significantly, with 81% agreement among the three indexers. Figure 3 shows the distribution of the classifications for the 2,428 abstracts where at least two indexers agreed on the main headings. This distribution confirmed our results from Figure 2, showing that some sections could be combined and others should be considered for subdivision. (For example, Fig. 3 shows that Section 9 probably would not contain enough abstracts over time to justify its existence, and so it should be combined with another related section.)

A number of studies have observed that indexer consistencies generally do not exceed 50%, and it has not changed significantly in the past 30 years (Leininger, 2000). Siert and Andrews (1991) found that 71 pairs of duplicate records in ISA had the same descriptors in about 48% of the cases. Reich and Biever (1991) compared the descriptors assigned by different indexers from the same thesaurus in two agriculture databases and found that they averaged between 24% and 45% agreement in different samples. Our results show significantly higher consistencies than those found previously—in 35% of the cases, all three of us agreed on the assignment, and in an additional 45% of the cases two of us agreed. Only 19% of the cases did all three of us disagree on the main heading assignment of an abstract.

It is of interest to note that in this work, we assigned headings based on the abstracts alone, in contrast to most
TABLE 1. Overall indexer consistency.

<table>
<thead>
<tr>
<th>Indexer agreement</th>
<th>Main and subheadings</th>
<th>Main headings only</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Number</td>
<td>% of total</td>
</tr>
<tr>
<td>All 3 Agree</td>
<td>793</td>
<td>26.4</td>
</tr>
<tr>
<td>A-B Agree</td>
<td>402</td>
<td>13.3</td>
</tr>
<tr>
<td>A-C Agree</td>
<td>424</td>
<td>14.1</td>
</tr>
<tr>
<td>B-C Agree</td>
<td>484</td>
<td>16.1</td>
</tr>
<tr>
<td>Only 2 Agree</td>
<td>1310</td>
<td>43.6</td>
</tr>
<tr>
<td>All Disagree</td>
<td>901</td>
<td>30.0</td>
</tr>
</tbody>
</table>

indexing studies, in which indexing is done from the full article. Despite indexing from the abstracts, the level of agreement among us is significantly higher than has generally been observed in the indexing studies cited above. Some possible reasons are:

(1) We all have significant subject expertise in information science, as well as many years of continuous practical experience in the field. (Two of us have each been working in various areas of information science for over 25 years, and the other one of us has been involved with abstracting and indexing for 12 years.)

(2) We had the benefit of previous exposure to the material used in this study. Because of our close involvement with and interest in ISA, this research was the second time we had seen either the abstracts or the articles they represented.

(3) As we proceeded through the corpus of material and questions arose, we learned from our experiences in choosing terms for the items indexed earlier in the process.

(4) Because we used professionally written abstracts in our tests, we could be confident that each abstract clearly identified the main concepts of the articles they represented. Because the abstracts were brief and focused, there was less uncertainty (and, hence, less disagreement) as might have occurred if we had used complete articles.

**Final Taxonomy Development and Use**

As we were proceeding through the first taxonomy validation, we began drafting revisions to correct deficiencies that were uncovered. Besides using our practical experience gained in the field and the experience of the validation test, we also examined the concepts propounded in Hawkins’ earlier article and ensured that they were incorporated into the taxonomy. A large part of this work was done empirically because ISA production continued as this research proceeded. We were thus able to take advantage of what authors in the field were publishing at that moment. In the “final” taxonomy (we regard it as final only with regard to this research project), the 13 main headings were consolidated into 11, and a number of the subcategories were realigned as well.

We also conducted a second validation test. Two of the authors manually indexed the May, June/July, and August 2001 issues of ISA, containing a total of 1,265 abstracts. The abstracts were copied from the master ISA production database to a separate Microsoft Access database, and the taxonomy was also converted to an Access database. A simple form, shown in Figure 4, displayed the title and abstract fields of each abstract and also provided an interface to the taxonomy, thus facilitating searching the taxonomy and entering the appropriate subject code. This second test resulted in the taxonomy presented in the Appendix, which became the “ISA Taxonomy” at the beginning of 2002. To date, six issues of ISA, containing a total of 2,692 abstracts have been produced using it. Each ISA abstract is required to be assigned to at least one subject classification, and it can be assigned to up to three subject classifications. For the 2,692 abstracts, a total of 3,504 subject classification assignments were made. Their distribution is shown in Figure 5.

Figure 5 shows that the abstracts are well distributed over the 11 main sections and subsections of the taxonomy. There are a number of heavily posted sections as well as a number with only a few postings. We examined each of the subsections shown below, all of which display a low number of postings, and concluded that they should remain as discrete subsections (Table 2).

Most of these subsections represent significant or growing areas of information science or, because of the focus of the field, they are important, even though not many articles concerning them are currently being published. We must remember that our sample for this test included only 3 months’ worth of data from the then most recent issues of ISA. In one or two cases, our years of experience in the field influenced our decision.

We also considered the three most highly posted subsections to determine if they should be subdivided (Table 3).

It is hardly surprising that these subsections are highly posted; because they represent the essence of the field, they contain many articles on closely related topics. We therefore decided not to subdivide them.

**Conclusion**

The authors of this research are an editor, a reference librarian, and an abstracter/indexer and thus represent three major communities within the information science field. We feel strongly that the involvement of representatives of these three user communities was critical to the success of this research because each one approaches use of a taxonomy from a different viewpoint:

(1) An editor views a taxonomy as a means of organizing a body of information and elucidating its hierarchical relationships, moving from the general to the more specific.

(2) The abstracter/indexer uses a taxonomy to help decide where to classify articles in the database. Indexers deal with the innuendos and subtleties of actually using the classification framework and can give appropriate feed-
FIG. 3. Main heading distribution—two indexers agreed.
At the heart of any discussion of the information sciences, at least when this discussion is held in the English language, lies the problem of the variant (and shifting) set of concepts and meanings of the terms "library" and "information." Outlines, for the English language only, some of the diverse meanings of information, and considers their consequences for the terminology of the information sciences. Focuses on the variant relationships between information and related concepts, particularly data and knowledge. Includes an account of the view of information taken in the hard and soft methodologies of systems science, as well as the new discipline of "information physics." From this, offers commentary on the changing meanings of complex terms such as "information technology," and "information literacy," as well as those complex terms involving management, information management, knowledge management, document management, etc. Gives a similar treatment to the terminology around the "library" concept, particularly in view of the change toward viewing a library as an organized virtual information space, rather than a physical environment.

FIG. 4. Indexer's taxonomy interface form.

FIG. 5. Classification distribution using new taxonomy.
TABLE 2. Subsections with few postings.

| 1.8 | Operations research/mathematics |
| 6.3 | Secondary publishing |
| 8.4 | Geographic information systems |
| 9.6 | News |
| 11.3 | Contracts and licensing |
| 11.4 | Liability issues |
| 11.6 | Information policies and studies |

TABLE 3. Highly posted subsections.

| 1.4 | Information retrieval research |
| 1.5 | User behavior and uses of information systems |
| 10.3 | Library automation, operations, and strategic planning |

back on its ease of use, inconsistencies, overlaps, and/or gaps in the structure. They have the unique skill of quickly identifying and summarizing the most pertinent and germane points made in an article, while usually working under significant time constraints.

(3) Reference librarians, online searchers, and researchers view taxonomies as entry points into databases and tend not to place complete reliance on them in their quest for information. Because such information seekers are either ultimate users of the information being sought or are acting in an intermediary mode on behalf of the user, they do not expect an editor or indexer to always know the concepts or special requirements of an information query as well as they do. They therefore view a taxonomy as only one component of their search strategy and incorporate additional synonymous terms or relationships that may apply to an information request. Despite their potentially disparate viewpoint, users must be included in the design and evaluation phases of development of the tools provided for their use. This principle is important regardless of what is being developed, and developers ignoring it take a significant risk that their products will not succeed in the marketplace.

The information science taxonomy we have developed represents a major improvement over the previous one. It reflects the modern-day definition of information science, has a practical orientation, and will be easy for ISA’s users to employ, both in reading the printed issues of the publication and in searching the electronic database. It is flexible and easily maintained; as technology advances and changes, new classifications can be easily added and obsolete ones deleted.

Although we are using the taxonomy that resulted from this work in the production of ISA, it has a much wider application. Our taxonomy is an information science taxonomy, and we suggest that it is applicable to all aspects of the discipline—one which is constantly changing and reinventing itself, and can be expected to do so in the foreseeable future—because it is based on the actual record of work in the field, as recorded in its publications. It is appropriate to emphasize again that

A&I publications...not only record the history of a discipline...but they can serve as "watchdogs" of changes. (Hawkins, 2001)

The taxonomy presented here uses the information science concepts, definition, and map previously developed and presents them in an organized hierarchical view of the field. Other limited classifications of information science exist, and were used as a starting point for this research, but as far as we know, the taxonomy we present here is the only one that directly addresses the field. As such, it makes a significant contribution.

Appendix. Information Science Taxonomy

1. INFORMATION SCIENCE RESEARCH

1.1 Basic concepts, definitions, theories, methodologies, and applications
1.2 Properties, needs, quality, and value of information
1.3 Statistics, measurement
Bibliometrics, citation analysis, scientometrics, informetrics
1.4 Information retrieval research
Searching techniques (Boolean, fuzzy, natural language), the search process, precision/relevance, ranking/recall, searching models, query formulation, inverted files, updating, database structures
1.5 User behavior and uses of information systems
Searcher tactics, information overload, user surveys, usability studies
1.6 Human-computer interface
Human factors, ergonomics, design issues

References

1.7 Communication
   Editing, writing, linguistics, Internet authoring and design principles
1.8 Operations research/mathematics
   Modeling, Boolean logic, coding, systems analysis, algorithms, compression
1.9 History of information science, biographies

2. KNOWLEDGE ORGANIZATION
2.1 Thesauri, authority lists
   Taxonomies, ontologies, semantic networks, nomenclatures, terminologies, vocabularies
2.2 Cataloging and classification
   Tagging, metatags, Dublin Core, DOIs, OPACs, MARC, AACR2, topic maps, cataloging processes and theories
2.3 Abstracting, indexing, reviewing
   Automatic indexing and abstracting
2.4 Standards and protocols
   NISO, Z39.5, XML, SGML, HTML, Open Archives Initiative (OAI), Encoded Archival Description (EAD), OpenURL, portable document format (PDF)

3. THE INFORMATION PROFESSION
3.1 Information professionals
   Intermediaries, searchers, reference librarians, information brokers, translators, educators, librarians and librarianship, mentoring, career outlook, future of the profession, professional ethics, skills and competencies
3.2 Organizations and societies

4. SOCIETAL ISSUES
4.1 Information ethics, plagiarism, credibility
4.2 Information literacy, lifelong learning
4.3 The Information Society
   Universal access and accessibility, technological and socioeconomic impacts of information, technology forecasts, information flows, futures scenarios, preservation

5. THE INFORMATION INDUSTRY
5.1 Information and knowledge management
   Knowledge transfer in organizations, business strategies
5.2 Markets and players
   Vendor profiles and interviews, trends
5.3 Economics and pricing
   Business models, value chain
5.4 Marketing, e-commerce

6. PUBLISHING AND DISTRIBUTION
6.1 Print
6.2 Electronic
   E-journals, e-books
6.3 Secondary publishing
   Abstracting and indexing services, directories
6.4 Scholarly communication
   Peer review process, future of journals, dissertations, grey literature

7. INFORMATION TECHNOLOGIES
7.1 Internet
   World Wide Web, Invisible Web, Deep Web, search engines, browsers, hypermedia, Listservs, bulletin boards, portals, gateways, directories, pathfinders
7.2 Intranets, Web conferencing
7.3 Software
   Programming languages, operating systems, platforms
7.4 Hardware
7.5 Multimedia
7.6 Document management
   Imaging, scanning, text retrieval, digitization, records management, bookmarking, hypertext systems, preservation technologies, digitization, linking and electronic cross referencing, storage, digital rights management
7.7 AI, expert systems, intelligent agents
   Cybernetics, visualization and mapping, data mining, pattern and character recognition, search agents and robots
7.8 Telecommunications
   Networks, wireless and satellite information delivery, Palm Pilots and other PDAs, LANs and WANs
7.9 Security, access control, authentication, encryption
   Digital watermarking
7.10 Other

8. ELECTRONIC INFORMATION SYSTEMS AND SERVICES
8.1 Information searching and retrieval systems and services
   Bibliographic, numeric, and image databases; descriptions of online services
8.2 Customized information systems, alerting, current awareness
8.3 Document delivery systems and services
Appendix. (continued)

8.4 Geographic information systems

9. SUBJECT-SPECIFIC SOURCES AND APPLICATIONS
9.1 Physical sciences
   Chemistry, physics, engineering, earth sciences, computer science, energy, mathematics
9.2 Life sciences
   Medicine, biosciences, agriculture, environment
9.3 Social sciences, humanities, history, linguistics
9.4 Business
   Management, economics, companies
9.5 Law, political science, government
   Patents and trademarks, intellectual property, case law
9.6 News
9.7 Education, library and information science, ready reference
9.8 Other/multidisciplinary
   Biography and genealogy databases, encyclopedias, databases of theses and dissertations

10. LIBRARIES AND LIBRARY SERVICES
10.1 Library descriptions and types
   Special, government, academic, and public libraries, archives, museums, state and national libraries, depository libraries
10.2 Library services
10.3 Library automation, operations, and strategic planning
10.4 Library consortia and networks, coalitions, cooperatives
10.5 Digital and virtual libraries, hybrid libraries
10.6 Education and training
   Distance learning, continuing education, bibliographic instruction library schools, courses and curricula

11. GOVERNMENT AND LEGAL INFORMATION AND ISSUES
11.1 Intellectual property protection
   Copyright issues and implications, fair use, trademarks, patent law
11.2 Legislation, laws, and regulations (except copyright)
11.3 Contracts and licensing
11.4 Liability issues
   Filtering, censorship, privacy
11.5 Sources of public information
11.6 Information policies and studies
   Security, encryption, privacy, freedom of information, censoring, national and other information policies
11.7 Systems and infrastructure
   Technology transfer