Title

Storage and search

Author

Buckland, Michael

Publication Date

2023-12-14

Michael K. Buckland, “Storage and Search” in *Information: A Historical Companion*, ed. by Ann Blair and others (Princeton University Press, 2021), 786-789. As submitted.

STORAGE AND SEARCH

All information history can be seen as the augmentation of direct communication (speech and gesture) with a relentless increase in indirect communication, either store-and-forward communication (letters, publications, email) or store-and-search communication (archives, libraries, personal notes, and now “big data”). The pervasiveness of this development has been accompanied by such wide differences in terminology that the essential identity of the underlying structure has been obscured.

Direct communication in the form of speech and gestures is transient, even if the effect may be lasting, but recordings and documents endure for a greater or shorter period and much of information history is concerned with the consequences. One consequence is that what endures will tend to have a more lasting influence than what is transient, provided that it is preserved and discoverable. The steady growth of information objects afforded by writing, printing, telecommunications, photography, and recording equipment using digital and other technologies has created and progressively exacerbated two challenges: how to store them and then how to find the items most suited for some purpose. The robustness and durability of records vary greatly by medium and circumstances. Repeated loss through disasters, such as the loss of the library of ancient Alexandria, and deterioration have gradually led to best practices for preservation and conservation. The long preeminence of the printed book in codex format meant that techniques for storage and search (notably bibliography and cataloging) have been narrowly focused on that medium. Work remains to be done generalizing these important techniques to newer media.

Ordered and Unordered Storage

As the number of items stored increases, the task of finding any particular one, or any that have desired characteristics, becomes progressively more difficult. In the simplest case of items stored at random a *serial search* will be more reliable and more efficient than searching randomly, but the average effort per search rises steeply as the collection size increases unless there is an ordered arrangement enabling one to go directly to a desired item at whatever location it occupies (*random access*). But a single ordering can support search only by the criterion used for that ordering and fails to support other preferences. For example, large libraries can economize on space needs by sorting books by size then shelving books of each size in the order received (*numerus currens*), but this is not helpful for readers seeking books on a topic.

Providing multiple copies for multiple sequences for multiple purposes is unaffordable, but the same effect can be achieved by adding virtual copies in any ordering desired (*index, inverted file*). Libraries, for example, typically shelve books in a classified subject arrangement and then provide additional virtual arrangements using as surrogate copies catalog records pointing to the book on the shelf. Traditionally, virtual arrangements by author, by title, and by verbal subject headings have been provided either as separate author, title, and subject catalogs or interfiled into a dictionary catalog for readers in addition to a shelf list for library staff. Similarly, bibliographies usually provide a single primary sequence augmented by multiple orderings provided via indexes. In a digital environment, the ease of generating multiple indexes

and of displaying parts or all of the stored items means that the primary ordering, typically a sequentially assigned identifier for administrative purposes, is irrelevant for the searcher. The relationship between the descriptive indexing (*metadata*) assigned to each document and the document itself is, in effect, inverted to constitute a crucial form of infrastructure whereby the systems of metadata are navigated to find the documents associated with each descriptive term.

Computer-based data sets are used the same way. A search command locates specified values in a serial search, with or without *inverted files* (indexes) for efficiency. Relational database management systems provide for multiple alternative virtual arrangements.

Dynamic Collections

Useful collections are typically dynamic as new items are added and in some cases old ones removed. Unless items are merely added in the order received, new items disrupt the ordering. Librarians leave some space on the shelves for expansion, but cannot predict reliably where expansion will be needed. Until the nineteenth century libraries used *fixed location*, whereby books were arranged by topic but assigned to a specific shelf location. The call number would specify the section of shelving, the shelf, and the position on the shelf. It did not require many additional books to disrupt this system. The decimal classification of Melvil Dewey solved this problem in two ways. First, books were shelved by classification number and so were ordered relative to each other (*relative location*) and not to any particular shelf. Moving a collection to different shelves does not affect the classification-based call numbers, nor therefore the shelving order. Second, since the classification system used a decimal notation, it is indefinitely hospitable to the insertion of new topics at any point at any time.

Scholars have long developed dynamic storage systems for their notes based on slips of paper or cards to allow flexibility for expansion and rearrangement for use or as a convenient way to prepare copy for printed bibliographies or other reference works.

Classification and Indexing

Arrangement by author, by title, and by date are relatively straightforward compared with arrangement by topic. Classified arrangements were long preferred to verbal subject headings and there was little use of alphabetizing beyond the initial letter until after printing developed in Europe. By the end of the eighteenth century reliance on scholar librarians and their knowledge of the literature, especially as represented in the local collection, was failing in the face of expanding knowledge, the relentless growth in publication and collections, and the increasing numbers of readers wanting to search and to discover by themselves. The technical development of librarianship by Martin Schrettinger, who coined the term “library science” for it in 1808, and others, was a deliberate response to this crisis.

Classification systems have three requirements: the ordering of concepts into a linear sequence; a form of notation; and, if verbal headings are not used, an index from natural language terms to that notation (e.g. Dewey’s “relativ” index; today a search term recommender).

Well-curated databases and other collections ordinarily have a carefully prepared and assigned set of categories. Nevertheless, effective and efficient searching depends on experience and familiarity with the categorization system used. The development of the internet increases access to a wider range of resources, but this inevitably introduces additional systems with unfamiliar categorization. Providing a mapping to terms in alien vocabularies from familiar

terminology is very expensive, does not scale well, and is inherently obsolescent. Statistical methods used in search-term recommender services can provide a cost-effective and easily updated alternative.

By the 1950s, very powerful and precise classifications (“indexing languages”) had been developed using *facet analysis* (e.g. using separate components for what, where, when, and who) and syntax (for distinguishing between, say, “man bites dog” and “dog bites man”), as well as ingenious mechanisms using edge-notched cards and optical coincidence (“peek-a-boo”) systems. But these powerful systems were hard to use and could not compete with the ease and economy of using punch cards and then digital computers to simply search text for any character string, with, later, some support for spell-checking and synonyms (vocabulary control) and specifying combinations of concepts at the time of search (*postcoordination*) which avoided the need to create these at the time of indexing (*precoordination*) and is more flexible.

Terminology

No single arrangement by subject can suit equally well all needs or the differing perspectives of specialists in diverse fields. Not only does each specialized community have its own particular use of words, but subjects and the relationships among them are always evolving. Similarly, subject terminology inherently faces obsolescence as both subjects and language itself change. The subject indexer, like Janus, must base terminology on established discourse (the past) in order to provide for future needs. Any assigned index term recedes fixed into the past while knowledge, language, and searchers evolve in new and unforeseeable ways.

Scale

The vast scale of contemporary digital resources (e.g. the web, Google Books, Amazon’s offerings, “the cloud”) and increasing ability to search for fragments within resources as a result of the move from printed to digital media, makes consistent human categorization impractical, so we have to fall back to basic search for character strings, augmented with elementary spell-checking prompts, mechanical surrogates for relevance (e.g., counting of citations or links), and, now, relentless inference from data collected from the searchers’ behavior.

Michael K. Buckland

Further Reading

Michael K. Buckland, *Information and Society*, 2017.
Ronald E. Day, *Indexing It All: The Subject in the Age of Documentation, Information, and Data*, 2014.
Daniel N. Joudrey and Arlene G. Taylor. *The Organization of Information*, 4th ed., 2017.
Markus Krajewski, *Paper Machines: About Cards and Catalogs, 1548-1929*, 2011.
Henry Petroski, *The book on the Bookshelf*, 1999.