Finding Information – The art of Indexing

We now live in an age where Google search is ubiquitous, and the 'find' function in Word and PDF documents is almost instantaneous. The challenge for most people is sorting through the long lists of information returned from a search to locate the most useful items; this was not always the case. As Dennis Duncan, a British writer, translator and lecturer set out in his book, *Index, A History of the: A Bookish Adventure*[1], the need for indexing first emerged in the 13th century and has been evolving ever since.

There are basically two indexing systems for books. The simplest is a listing of the important words that occur in a reference, identifying the pages or sections in which the word is used. The more complex system is built around topics and identifies the section of a book in which the topic is discussed, often indexing multiple publications.

Both systems were developed around the year 1230[2], and mark the change from a time when books were a valued artifact to be read and enjoyed, to one where books became an information repository to be used as a resource. In the 13th century very few people could read, and books were scarce, making the oral delivery of information vital either as a lecture or a sermon. But delivering a lecture (or sermon), required information to be sourced, organised, synthesised, and written down in preparation for the delivery. This means the presenter needed to use books – not just read books, but to be able to go back and use the contents of books as an information resource. The invention of the printing press would not occur for another 200 years (1440), so in 1230 books were still an incredibly valuable resource in limited supply.

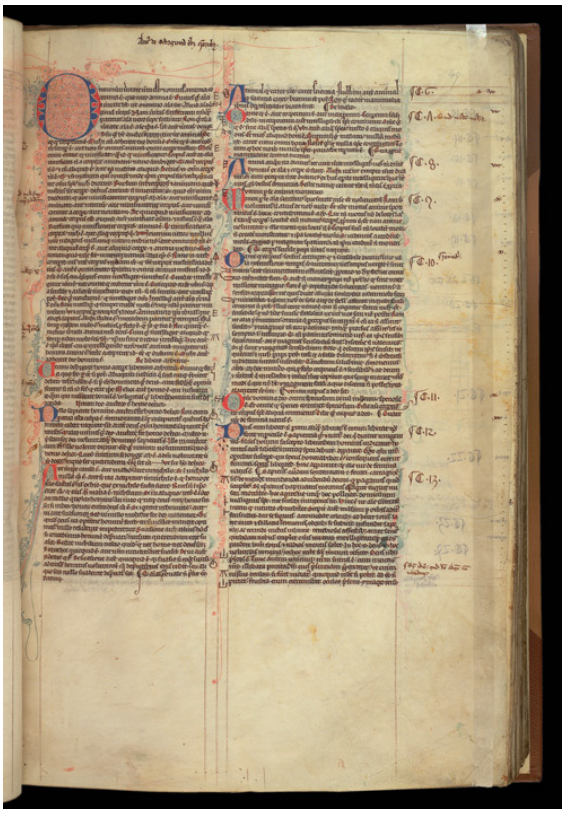

Figure 1: Source British Library - On Line Exhibition

The word index was invented in Paris, by a Dominican Abbot named Hugh of Saint-Cher. The Dominicans are a preaching or mendicant religious order, founded in 1216. Their calling was to have Friars live among the people in big cities and preach sermons to stop the flock from going astray.

To help his Friars write their sermons, Saint-Cher instructed a group at the Dominican Friary of Saint-Jacques, to create a word index, or a concordance, of the Bible. Every single word in the Bible was put in alphabetical order with a locator indicating where that word appears. The friars listed about 10,000 individual words and 129,000 locations. As a consequence of this

1 *Index, A History of the: A Bookish Adventure*, Duncan, D. Allen Lane, UK, 2021. ISBN: 9780241374238

2 To see the events discussed in this paper in a comprehensive historical timeline download ***Project Management - A Historical Timeline***: https://mosaicprojects.com.au/PDF_Papers/P212_Historical_Timeline.pdf

work (that still exists), the preaching Friars writing new sermons were able to find the information they needed reliably and consistently.

A parallel driver for indexing was the creation of universities, with Oxford being one of the earliest. Robert Grosseteste, a medieval English scholastic philosopher, taught at Oxford until his appointment as Bishop of Lincoln in 1235. Grosseteste read widely, and to help locate materials for his lectures, invented an indexing system based on symbols made up of curved and straight lines, circles, E-shapes, etc., which were added as annotations in each of his books. Different symbols represented different subjects, and in a separate general index he kept a record of where they were located. The result was a kind of parchment Google – once he's read and annotated a book, he knew where information on a subject was for future reference. This type of index is still called a general index.

With the advent of indexing, engaging with a book transitioned from being a linear process where the reader had all the time in the world to journey from end to end, to one where books became seen as storehouses of morsels of information. The indexes allowed people to use and research their books more efficiently, enabling them to preach or lecture at short notice.

800 years later, these concepts are still evolving. Unfortunately, the traditional concept of indexing is rapidly disappearing, the fundamental requirement for an index is a page number, and e-books don't have set pages, the page a word appears on changes depending on the font size and screen size selected by the reader. This is a pity, creating a good index is both an art and a craft, requiring interpretation and judgement to look at each passage and decide what words a person would use to look for to find that specific section of text.

On the other hand, Robert Grosseteste's concept of the general (or subject) index has moved from the world of academia to mainstream. Google indexes millions of pages of new information every day. Both Google and the various feeds to your PDA index then select what you see based on the topics you are interested in, filtered by the application of a liberal dose of artificial intelligence (AI).

The challenge for everyone in the modern era is being able to filter and validate the thousands of returns from a typical Google search while at the same time making sure their feeds are not too limited. The various systems will order the information you see in a way its AI systems calculate will give you the best experience. But *best* from system's perspective is that you like the result and will therefore use it again. This is not the same as offering the most accurate selection of information, particularly if there are contradictory viewpoints.

Library classifications

The concept of a general index can be enhanced by creating a taxonomy (or taxonomical classification). A taxonomy is a scheme of classification, especially a hierarchical classification, in which things are organized into groups or types. Among other things, a taxonomy can be used to organize and index knowledge (stored in books, documents, articles, videos, etc.). Examples include library classification systems[3], or a

[3] For more on the development of ***the first taxonomy for books, and a library catalogue system*** see: https://mosaicprojects.wordpress.com/2023/01/30/the-great-library-of-alexandria-the-first-google/

search engine taxonomy. Both are designed so that users can find the information they are searching for more easily. The ***PMKI Taxonomy***[4] created for the Mosaic website is a useful example.

The standard classification system used by most libraries today is the Dewey Decimal Classification (DDC)[5]. This is a proprietary system which allows new books to be added to a library in their appropriate location based on subject. The DDC was first published by Melvil Dewey, in the United States, in 1876. Its decimal number classification introduced the concepts of relative location and relative index. Numbers are flexible to the degree that they can be expanded in linear fashion to cover special aspects of general subjects[6]. For example:

500 Natural sciences and mathematics

510 Mathematics

516 Geometry

516.3 Analytic geometries

516.37 Metric differential geometries

516.375 Finsler geometry

A library assigns a classification number that unambiguously locates a particular volume in a position relative to other books in the library, on the basis of its subject. This number makes it possible to find any book and to return it to its proper place on the library shelves. The problem with the system is the size of the full classification index which grew to some 2000 pages. To overcome this problem the system has evolved into three versions:

1. The **DDC** standard edition is designed for the majority of general libraries by not attempting to satisfy the needs of the very largest or of special libraries. It is around 700 pages in length.
2. **Library of Congress Classification** having gained acceptance by most large research libraries, has become the full classification, and
3. The **Universal Decimal Classification** which is used for the organization of bibliographies rather than of books on the shelf.

The Dewey system was not the first, or only classification system. The *Cutter Expansive Classification* was published in 1882[7]. This classification system provided the basis for the Library of Congress top-level classification system. However, prior to and contemporary with these systems libraries classified documents according to their needs and (often) whims of the people in charge, however the 'do it yourself' approach is declining in general and research libraries (but increasing in organizations knowledge management systems and many web based search engines – see below).

Two of the oldest known classification systems are the seventh century BCE, system used in the royal Library of Ashurbanipal at Nineveh. It held 30,000 clay tablets, in several languages, organized according to shape and separated by content. The other is the *Pinakes*[8]. It was a catalogue of the bins that held the

---

[4] Download the ***Project Management Knowledge Index (PMKI) Taxonomy***: https://mosaicprojects.com.au/PDF-Gen/PMKI_Index.pdf

[5] See: https://en.wikipedia.org/wiki/Dewey_Decimal_Classification

[6] Note: There are additional features used in the DDC defined in a series of tables.

[7] See: https://en.wikipedia.org/wiki/Cutter_Expansive_Classification

[8] For more on the development of the ***Pinakes*** see: https://mosaicprojects.wordpress.com/2023/01/30/the-great-library-of-alexandria-the-first-google/

papyrus scrolls in the Library of Alexandra created during the third century BCE. The way the Pinakes catalogue was structured influenced the structure of similar listings in the West for centuries.

Finding the knowledge

The ever-increasing volume of information contained in books and other documents makes indexing and searching challenging. Literally millions of new reference books and papers are published annually. Finding what matters efficiently requires selected subsets of information within a knowledge management system[9]. The ***PMKI Taxonomy***[10] was created to make the information on the Mosaic website accessible. Some other excellent search engines in the world that specialize in books, science, other smart information include:

- www.refseek.com - Academic Resource Search. More than a billion sources: encyclopedia, monographies, magazines.
- www.worldcat.org - A search for the contents of 20 thousand worldwide libraries. Find out where lies the nearest rare book you need.
- https://link.springer.com - Access to more than 10 million scientific documents: books, articles, research protocols.
- www.science.gov - An American state search engine on 2200+ scientific sites. More than 200 million articles are indexed.
- www.pdfdrive.com - The largest website for free download of books in PDF format. Claiming over 225 million names.
- www.base-search.net - One of the most powerful research engines on academic studies texts. More than 100 million scientific documents, 70% of them are free

The challenge in designing a knowledge management system is to craft a system that contains accurate information, is easy to use and can adapt over time.

---

[9] For more on ***knowledge management systems*** see: https://mosaicprojects.com.au/PMKI-PBK-010.php#Process3

[10] Download the ***Project Management Knowledge Index (PMKI) Taxonomy***: https://mosaicprojects.com.au/PDF-Gen/PMKI_Index.pdf

First Published 3rd August 2022 – Augmented and Updated