Digitization breaks a document into units of which each is assigned a numbered position in the sequence of the document. From this perspective digitization can be viewed as a total indexation of the document. It is converted into units rendered for machine operations. This sequentiality is made explicit, by means of an underlying index.

Sequences and chains are orders of one dimension. Their one-dimensional ordering allows addressability of each element and [random] access. [Jumps] between [random] addresses are still sequential, processing elements one at a time.

(K) The index

Summa confessorum [1297-98], 1310. [7] (http://www.bl.uk/onlinegallery/onlineex/illmanus/roymanucoll/j/011roy000008g11u00002000.html)

[The] sequencing not only weaves words into statements but activates other temporalities, and *presents occurrences of words from past statements*. As now when I am saying the word *utterance*, each time there surface contexts in which I have used it earlier.

A long quote from Frederick G. Kilgour, *The Evolution of the Book*, 1998, pp 76-77:

> "A century of invention of various types of indexes and reference tools preceded the advent of the first subject index to a specific book, which occurred in the last years of the thirteenth century. The first subject indexes were "distinctions," collections of "various figurative or symbolic meanings of a noun found in the scriptures" that "are the earliest of all alphabetical tools aside from dictionaries." (Richard and Mary Rouse supply an example: "Horse = Preacher. Job 39: 'Hast thou given the horse strength, or encircled his neck with whinning?')
>
> [Concordance] By the end of the third decade of the thirteenth century Hugh de Saint-Cher had produced the first word concordance. It was a simple word index of the Bible, with every location of each word listed by [its position in the Bible specified by book, chapter, and letter indicating part of the chapter]. Hugh organized several dozen men, assigning to each man an initial letter to search; for example,

> the man assigned M was to go through the entire Bible, list each word beginning with M and give its location. As it was soon perceived that this original reference work would be even more useful if words were cited in context, a second concordance was produced, with each word in lengthy context, but it proved to be unwieldy. [Soon] a third version was produced, with words in contexts of four to seven words, the model for biblical concordances ever since.
>
> [Subject index] The subject index, also an innovation of the thirteenth century, evolved over the same period as did the concordance. Most of the early topical indexes were designed for writing sermons; some were organized, while others were apparently sequential without any arrangement. By midcentury the entries were in alphabetical order, except for a few in some classified arrangement. Until the end of the century these alphabetical reference works indexed a small group of books. Finally John of Freiburg added an alphabetical subject index to his own book, *Summa Confessorum* (1297—1298). As the Rouses have put it, 'By the end of the [13]th century the practical utility of the subject index is taken for granted by the literate West, no longer solely as an aid for preachers, but also in the disciplines of theology, philosophy, and both kinds of law.'"

In one sense neither subject-index nor concordane are indexes, they are words or group of words selected according to given criteria from the body of the text, each accompanied with a list of identifiers. These identifiers are elements of an index, whether they represent a page, chapter, column, or other [kind of] block of text. Every identifier is an unique *address*.

The index is thus an ordering of a sequence by means of associating its elements with a set of symbols, when each element is given unique combination of symbols. Different sizes of sets yield different number of variations. Symbol sets such as an alphabet, arabic numerals, roman numerals, and binary digits have different proportions between the length of a string of symbols and the number of possible variations it can contain. Thus two symbols of English alphabet can store 26^2 various values, of arabic numerals 10^2, of roman numberals 8^2 and of binary digits 2^2.

Indexation is segmentation, a breaking into segments. From as early as the 13th century the index such as that of sections has served as enabler of search. The more [detailed] indexation the more precise search results it enables.

The subject-index and concordance are tables of search results. There is a direct lineage from the 13th-century biblical concordances and the birth of computational linguistic analysis, they were both initiated and realised by priests.

During the World War II, Jesuit Father Roberto Busa began to look for machines for the automation of the linguistic analysis of the 11 million-word Latin corpus of Thomas Aquinas and related authors.

Working on his Ph.D. thesis on the concept of *praesens* in Aquinas he realised two things:

> "I realized first that a philological and lexicographical inquiry into the verbal system of an author has t o precede and prepare for a doctrinal interpretation of his works. Each writer expresses his conceptual system in and through his verbal system, with the consequence that the reader who masters this verbal system, using his own conceptual system, has to get an insight into the writer's conceptual system. The reader should not simply attach t o the words he reads the significance they have in his mind, but should try t o find out what significance they had in the writer's mind. Second, I realized that all functional or grammatical words (which in my mind are not 'empty' at all but philosophically rich) manifest the deepest logic of being which generates the basic structures of human discourse. It is .this basic logic that allows the transfer from what the words mean today t o what they meant to the writer.

> In the works of every philosopher there are two philosophies: the one which he consciously intends to express and the one he actually uses to express it. The structure of each sentence implies in itself some philosophical assumptions and truths. In this light, one can legitimately criticize a philosopher only when these two philosophies are in contradiction." [11] (http://www.alice.id.tue.nl/references/busa-1980.pdf)

Collaborating with the IBM in New York from 1949, the work, a concordance of all the words of Thomas Aquinas, was finally published in the 1970s in 56 printed volumes (a version is online since 2005 [12] (http://www.corpusthomisticum.org/it/index.age)). Besides that, an electronic lexicon for automatic lemmatization of Latin words was created by a team of ten priests in the scope of two years (in two phases: grouping all the forms of an inflected word under their lemma, and coding the morphological categories of each form and lemma), containing 150,000 forms [13] (http://www.alice.id.tue.nl/references/busa-1980.pdf#page=4). Father Busa has been dubbed the father of humanities computing and recently also of digital humanities.

The subject-index has a crucial role in the printed book. It is the only means for search the book offers. Subjects composing an index can be selected according to a classification scheme (specific to a field of an inquiry), for example as elements of a certain degree (with a given minimum number of subclasses).

Its role seemingly vanishes in the digital text. But it can be easily transformed. Besides serving as a table of pre-searched results the subject-index also gives a distinct idea about content of the book. Two patterns give us a clue: numbers of occurrences of selected words give subjects weights, while words that seem specific to the book outweights other even if they don't occur very often. A selection of these words then serves as a descriptor of the whole text, and can be thought of as a specific kind of 'tags'.

This process was formalized in a mathematical function in the 1970s, thanks to a formula by Karen Spärck Jones which she entitled 'inverse document frequency' (IDF), or in other words, "term specificity". It is measured as a proportion of texts in the corpus where the word appears at least once to the total number of texts. When multiplied by the frequency of the word *in* the text (divided by the maximum frequency of any word in the text), we get *term frequency-inverse document frequency* (tf-idf). In this way we can get an automated list of subjects which are particular in the text when compared to a group of texts.

We came to learn it by practice of searching the web. It is a mechanism not dissimilar to thought process involved in retrieving particular information online. And search engines have it built in their indexing algorithms as well.

There is a paper proposing attaching words generated by tf-idf to the hyperlinks when referring websites [14] (http://bscit.berkeley.edu/cgi-bin/pl_dochome?query_src=&format=html&collection=Wilensky_papers&id=3&show_doc=yes). This would enable finding the referred content even after the link is dead. Hyperlinks in references in the paper use this feature and it can be easily tested: [15] (http://www.cs.berkeley.edu/~phelps/papers/dissertation-abstract.html?lexical-signature=notemarks+multivalent+semantically+franca+stylized).

There is another measure, cosine similarity, which takes tf-idf further and can be applied for clustering texts according to similarities in their specificity. This might be interesting as a feature for digital libraries, or even a way of organising library bottom-up into novel categories, new discourses could emerge. Or as an aid for researchers to sort through texts, or even for editors as an aid in producing interesting anthologies.