Topic maps and the essence of indexing

Michel Biezunski

The topic maps model was invented to incorporate indexes with emerging technologies. But it goes beyond simply capturing the way indexes are published, and drills to the core of indexing. Michel Biezunski argues that human indexers still have much to offer in the realm of the World Wide Web and beyond, across digitized information repositories.

Introduction

At the beginning there was text. When typesetting moved from lead characters to phototypesetting using computers, the special characters used to produce typographical effects ('markup') were standardized, but instead of simply devising a common way to switch between font variations, such as italic or bold face, an intermediate step was created. This step gave publishers the freedom to express why they needed the variations. For example, was italic being used to distinguish a work or to mark a quotation?

Mapping was done at the tool level by formatting engines to convert a semantic distinction into a visual artifact. This principle of the separation between content and presentation was the foundation for the Standardized Generalized Markup Language (SGML) which was released in 1986.[1] At that time, computers were not only used by big corporations. Personal computers were adopted by many, mainly for the purpose of organizing their own data using spreadsheets, databases and word processing software. The concept of WYSIWYG (what you see is what you get) was the main force driving the market. People wanted to see on their screens exactly what they needed. This took publishers in two directions. On the one hand, the availability of WYSIWYG for creating text (word processing) and for layout (desktop publishing) meant that they gained the possibility of bypassing the process of sending a typewritten manuscript to a typesetting facility. On the other hand, organizations and corporations started to use SGML and later XML to organize their content according to a structure that complies with industry standards.

What became known as the topic maps model was thus designed from within the SGML community in the early 1990s as a way to make indexes 'interoperable'.[2] It was a response to the needs of Unix software vendors who were struggling to improve the consistency between the different flavors and providers. They also needed to respond to complaints about vocabulary changes when switching between brands. The topic maps model, was designed as a way to describe all kinds of navigational aids, encompassing not only indexes, but also thesauri, taxonomies, classifications, catalogs, glossaries, dictionaries, as well as tables of content and cross-references.

We were not the only ones with that goal. Soon after HTML took off and the world adopted the Web as its information platform, the World Wide Web consortium started to design an approach to encode metadata and express it into a graph. The community of librarians came together with the technologists. In 1995, in Dublin, Ohio, a meeting was held at the initiative of the Online Computer Library Center (OCLC), and produced what became known as the 'Dublin Core', which is now in use by many libraries throughout the world. The underlying language for metadata representation is the Resource Description Framework (RDF), in which information is represented as a graph of interconnected Web locations. This served a few years later as the foundation of the 'Semantic Web'.[3]

As the World Wide Web grows and books lose their unique role in spreading information, indexes (which are inseparable from the books that contain them) are now facing competition from other search techniques. Manual work is being contrasted with automated search algorithms. However, the distinction between automated search and manually produced indexes is somewhat misleading, because the companies providing search technologies employ many humans to refine the way search is conducted, and human indexers rely on tools to help them build indexes. The question remains about how to handle the 'last mile'. For example, online shopping helps customers to buy goods and appears as purely virtual as opposed to going to a store, but its side effect has been to multiply the amount of trucks on the roads to carry the goods from warehouses to the customers. Similarly, search can be done without having to go to a library or buy printed materials, but there are still many people employed to increase the accuracy of search engines.

What do topic maps have to do with indexes?

From a typesetting perspective, indexes are a part of a book containing an alphabetic list of terms, followed by numbers, separated by commas, or dashes to express ranges. Indented lines indicate to the reader that more specific information relevant to the term is available. Some terms are not followed by page numbers, but by the word 'see' (frequently in italics) to refer to another term. The 'see also' indicator is added to a heading or a 'sub-heading' to indicate that another term is also relevant in this context.

From a computerized perspective, structured markup can help capture the various components of each index entry. Many XML schemas have provisions for indexes that describe the granularity of components used. Indexing software makes these structural components user-friendly, so that the creators of the index can use form-like menus to enter the information without having to worry about

entering the actual markup. Software tools also enable indexes to be embedded into the document content, and the page numbers are computed at rendition time.

For example, the Docbook architecture, based on XML (previously SGML) is a good example of structured, semantic markup, which makes the components of an index entry explicit:

```
<indexterm>
<primary>information</primary>
 <seealso>data</seealso>
 <secondary>dissemination</secondary>
</indexterm>
```

These approaches respect the classic notion of an index.

The topic maps concept, in contrast, appears more disruptive. Traditional well-known navigational aids, such as indexes, tables of content, cross-references, thesauri, and taxonomies are pre-resolved queries in a topic database. They do not need to be stated explicitly because every user is expected to know what they mean. For example, an index is an alphabetically organized list of names that represent specific units of meaning together with locators enabling navigation to the resource that is relevant to that unit of meaning. Sometimes, these units of meaning are connected to others.

Topics as abstractions

Humans build mental representations of abstract concepts or things. When we communicate, we use words to express those meanings. The words we use depend on the language we speak, the education level we have, and the people we are talking to. Even then, the same word can express different meanings, and one meaning can be expressed with different words. Any subject of conversation has a meaning. We can talk about anything, including the words themselves (How do you spell 'organization'?) or a subject that is different from the thing that it describes (What is this organization doing?). The categories in which things are classified are also subjects for discussion. In the Topic Maps paradigm, a 'topic' is a subject that occupies an abstract location in an abstract space. The word 'topic' has been chosen because it refers to a subject, and it comes from the Greek work topos, meaning 'place'.

The fact that each distinct meaning occupies a specific place in a topic space is conceptually abstract, and we decided to give it a computer representation as an 'object', that is, as data stored on a computer with its associated methods. We also added the ability for any topic to be connected to any other. And we named the graph of topics a 'topic map', since it was primarily aimed at navigational purposes.

Properties of topics

Then, our next step was to design properties for these topics. These properties are intended to facilitate ways of navigating information resources, and capturing the features vastly used by the traditional navigational aids. A topic can have an unspecified number of names. The topic is said to 'occur' on the specified locations in the content that are considered to be relevant. In other words, the location is described as an occurrence of the topic. A topic can be related to others through associations whose semantics are user-customizable.[4]

Other properties of topics are proposed as short circuits. For example, a type is a property of a topic that describes a higher-level category that can be used as a filter. It is also the basis for building hierarchical taxonomies. Alternatively, a topic type can also be described as just another topic associated with the original one through an association with the semantic of 'typing'.

As names are not used as topic identifiers, the same name can be used in two different topics. For example, the name 'New York' can be used for a topic representing that city and for another topic representing that state. Names are disambiguated using a property called 'scope'. One topic named New York can be assigned the scope 'city', and the other one 'state'. This is not the same as assigning types to them. In an early iteration of the Topic Maps model, we thought that types could be used for disambiguation. Sure enough, the topic New York that has the type 'City' is different from the topic New York that has the type 'State'. But consider the '14th Street' subway station in New York, as there are several of those – one is in the scope '6th Ave' while another is in the scope '7th Ave'. Here it would be a stretch to consider that a subway station belongs to the type '7th Ave'. Both of those topics could instead reasonably be assigned the same type 'subway station', while being differentiated by their scope property instead.

An example of a frequently used topic without a name is a cross-reference included in the text of a book. An author can decide to write 'for more details, see chapter xxx'. There is no explicit topic here, but the author implies that discussion of the subject continues in another location. The topic does exist and has two occurrences but no name. This is a situation that is perfectly acceptable in a topic map context.

Occurrences of topics

In topic map terms, an occurrence is a relation between a topic and a location. In printed materials, locations are usually indicated by page numbers, but they could also be section numbers or chapter numbers. When content is digitized, a location can be expressed by an invisible identifier. Locations can be expressed as the starting point, by a whole element (for example, a chapter can be a whole element), as a URL for a web page, or as a specific fragment of a page. Languages such as XPath also enable locations to be expressed as a result of processing. For example, a location could be expressed as 'the second paragraph of the third chapter'. An occurrence can itself have type(s) that are scoped in order to provide more context to motivate a user to select one among others.

Topics can be connected to each other by relations which can optionally indicate a given semantic. The relations are not limited to hierarchies. These relations, called 'associations' in the topic maps model, may also have types, and the role that every participant plays in the relation can also be expressed. A topic can be related to other topics through the same or different kinds of relations. Altogether, the topics are connected through a graph. In general, there is

no expectation that a topic must have a 'subtopic': a hierarchical relation is one among several possible relationships between one topic and another. In an index, when entries have subentries, it is tempting for non-professional indexers to consider that the subentries are 'included' into the entries. This is visually the case, but not necessarily semantically. The two terms may be at the same semantic level, when related through 'and', as in 'Affirmative action' with the subheading 'and faculty hiring'. Or the inverse relationship can be used within the same index, by flipping the entry with its subentries. Intellectually speaking, it is possible to regard every subentry as a plain topic in its own right and 'entry-subentry' as one kind of relation that is slightly different from the 'see also' relationship, but not so much after all.

Expressing an index as a topic map

Table 1 illustrates a way to create a topic map from an index. There is more than one way to express an index as a topic map; for example, the semantics of associations between topics can be further detailed into multiple association types.

Beyond indexes

Now consider a classical book index within a topic map context. Topics are subjects that have been isolated and to which names have been given. Locations are pages, and the location indicators are page numbers. Each topic gets occurrences: that is, the page numbers or ranges in which it appears. Then various types of relations are declared for use between topics. For example, one can be called 'closely related to' and will end expressed as subentries. Another is the 'see also' relationship, and a third is the 'see' relation, which could alternatively be represented by adding a name to a given topic.

Once the topic is built, it becomes possible to issue a query that contains an alphabetical list of all topic names, followed by their location indicators, followed by an indented block containing closely related topics, and in which the 'see' and 'see also' relations appear. This is no more, no less, than an index. If the ending point is to come back to a traditional index, what have we gained? Looking under the hood on what an index is (semantically) as opposed to how it looks (visually) helps integrate indexes into a broader context. A similar process can be used to describe a variety of other navigational aids, including taxonomies, classifications, product catalogs, tables of authorities, concordance tables, databases, tables of content, ontologies, glossaries, dictionaries, encyclopedias, and spreadsheets. By providing a common platform it becomes possible to design hybrid products that include master indexes, indexes with glosses or other features.

Scaling topic maps

The traditional topic map tools have been focused either on providing custom processing for specific information repositories or on merging topic maps. The appetite for

Table 1 Indexes and topic maps

Index components	Corresponding topic map constructs
Index entry	Topic
Index term	Topic name
Page number	Occurrence locator
Page number range	Occurrence locator
Set of page numbers following a heading	Occurrences
Subheading	Association to another topic, with varying association type. Generally speaking, the association type can be described as "closely related to". More specific association semantic can be explicitly declared depending on the indexer's intent.
Subheading term	Topic name. If considered a topic per se, the name of the topic may concatenate the heading and subheading.
See also	Association to another topic, with an association type 'Somewhat related to'
See	Either considered a synonym, i.e. adding the target of 'see' to the topic name, or association to another topic with an association type 'Also related to.' The topic at the origin of the 'see' relation has no occurrence.
Multiple indexes (e.g. index of names, concepts, locations)	Each topic in every index is assigned a corresponding topic type (e.g. 'Person,' 'Concept,' "Location')
Heading/subheading flipping	Associations between those topics can be declared as 'bidirectional' or 'unidirectional.' Bidirectionality means that flipping is enabled.

merging topic maps did not find a market sufficient to stabilize the products, and therefore there are not many topic map software products available. But there are a number of products which have similar functionalities, without referring explicitly to topic mapping. They are often built as applications of graph databases.

There is still an important potential in the future for providing a connecting tissue to break the traditional barriers between different navigational aids. But this has to be measured against the scale of the information covered. Over-arching taxonomies and topic descriptions, supposedly valid over the whole Web, such as Wikidata, are trying to achieve this goal. But the more information is linked to a particular URL representing a subject, the less flavor it has. There is a threshold above which too much information is detrimental. When getting a million or more hits from a Google search, most people will not continue browsing past the first few pages of results. And since we do not know why Google is giving us the results we get in this order, it is hard to know whether we have actually found what we need. This problem does not have an easy fix.[5]

However, Google itself provides topic map navigation. For example, when you search for San Francisco, a 'knowledge card' appears on the screen with a map, a photograph, and some data about San Francisco, including the current weather and the name of the mayor. There are also related topics ('California', 'USA', 'Los Angeles', 'San Francisco Area', 'New York City'), which may perhaps appear differently depending on what Google knows about each of us. This information is based on a topic map, called the Knowledge Graph, which was acquired by Google from a company called Metaweb, which explicitly built its product, Freebase, as an implementation of a topic map. This particular topic map may not appear particularly useful, because it has not been designed with any particular usage in mind. Similarly, the integrated taxonomies which cover a lot of territory are by nature not focused on particular interests. All information is therefore not made to be merged. Another issue of a controlled vocabulary is that it may prevent certain original or rare concepts ever to be seen. By contrast, an indexer of a book is free to choose terms to express the specific subjects that the authors have developed in their book.[6]

Conclusion

At the end of the day, consumers of information products need to find relevant information resources, and a company that is in the business of publishing information must make sure that consumers will find their resources. This information needs to be visible, and therefore needs to be connected. But it also needs to be unique. Finding how to articulate these seemingly contradictory requirements is difficult, and there is no simple answer. This is where indexers, with their accumulated knowledge and expertise in analyzing information, have much to offer. Even with breakthroughs in artificial intelligence and machine learning, humans remain the most valuable resource when things become tricky. Indexers tend to like to address and overcome challenging situations. Even if the products they will end up contributing to do not visually resemble back-of-the-book indexes, the deep nature of the work that they are doing will remain necessary – even indispensable.

Notes

1 SGML became an inspiration for various developments. A structured music notation language was developed and was later generalized into an approach integrating hypermedia and time/based information description (HyTime). More significantly, an SGML-like language, HTML, was developed with a minimal markup to encode documents that could be visualized on computers and connected to each other via hyperlinks. The availability of HTML sparked the rapid emergence of the World Wide Web (Web). Because SGML was too complicated to be handled by Web browsers, a simplified version called XML was designed. XML removed the technical refinements that were rarely used, in the hope that it could become the lingua franca of the Web. It did not turn out that way, but XML has been massively adopted as a data interchange format (Kasdorf, 2004). Ironically, XML is now considered overwhelming, and JSON, the Javascript Object Notation, is the most universally adopted format for transferring data to various contexts, including the Web.

2 For more details about the history of topic maps, see Northedge (2008).

3 One of the early applications of the Semantic Web was known as 'Friend of a Friend', a vocabulary used to demonstrate the power of interconnecting people. This was the precursor of Facebook. In the early 2000s, the amount of information available on the Web increased in such proportion that powerful search technologies were developed, such as Google, which became the leader in the industry. The tension between the pressure organized by the public sector, imposing compliance to standards to its contractors, and the privatization of information, including personal information, which was the quid pro quo for companies such as Google or Facebook to provide their services for free, turned clearly in favor of the private sector. For example, Google has digitized millions of books in cooperation with libraries. The scanned texts were deposited not only in Google Books, but also in the HathiTrust. Google acted in the public interest and at the same time increased its value as a knowledge company.

4 A well-known article on topic maps focuses on topics, associations and occurrences (TAO): Pepper (n.d.).

5 Discussions are starting about who owns knowledge, what are the credentials requested for accessing it, and whether it is possible to regulate or to legislate about it. This includes the ongoing discussions on net neutrality and privacy protection rights. Simplistic answers to these complex questions also exist; for example, attempts made by authoritarian regimes to control the flow of information made available to their citizens.

6 See Johncocks (2005) for an earlier discussion of these issues at a technical level. People who work on content from an intellectual perspective (and that includes indexers) may have something interesting to say about what we can do as a society to address this new, emerging issue of not only how to deal with too much information, but also how to account for the selection of what information is being shown.

References

Cournington, C. (2010) 'Topic maps and indexing: greater, more cost-effective search access.' *Key Words* **18**(4), 132–8.

Johncocks, B. (2005) 'The myth of the reusable index.' *The Indexer* **24**(4), 213–17.

Kasdorf, W. E. (2004) 'Indexers and XML: an overview of the opportunities.' *The Indexer* **24**(2), 75–8.

Northedge, R. (2008) 'The medium is not the message: topic maps and the separation of presentation and content in indexes.' *The Indexer* **26**(2), 60–4.

Park, J. and Hunting, S., (eds.) (2003) XML *Topic Maps*. Boston, Mass.: Addison Wesley, Pearson Education.

Pepper, S. (n.d.) 'The TAO of topic maps – finding the way in the age of infoglut.' https://ontopia.net/topicmaps/materials/tao.html

Michel Biezunski is an innovator, providing consulting and application services for publishers. His focus is on enabling his customers to curate high value information content, and he designs and creates web-based applications for that purpose. Michel has been working mostly with government agencies and academic publishers. He has been actively involved in creating, publishing and using the Topic Maps standard. His background is in history and philosophy of science. He is the founder of Infoloom, a company based in New York City. Email: mb@infoloom.com