Patterns of Information Classification

FILIPE FIGUEIREDO CORREIA, Departamento de Engenharia Informática, Faculdade de Engenharia, Universidade do Porto
ADEMAR AGUIAR, INESC Porto, Departamento de Engenharia Informática, Faculdade de Engenharia, Universidade do Porto

Providing efficient access to information can be approached in different ways, but ultimately implies the creation of an INDEX, represented with an indexing language, like a TAXONOMY, a THESAURUS, an ONTOLOGY or a FOLKSONOMY, to name a few. Each of these languages strikes a different balance between the effort to create and maintain the index, the effectiveness of knowledge capture, the guidance that readers can get, and how efficiently they can get it. Furthermore, in a world in which more and more information is available, two issues gain particular importance in the creation of an index: how can it be done collaborative, and how can the index abstract and express information more richly.

Categories and Subject Descriptors: D.2.11 [**Software Architectures**] Patterns; H.3.1 [**Information Storage and Retrieval**] Content Analysis and Indexing

General Terms: information, classification

Additional Key Words and Phrases: indexing, index, taxonomy, thesaurus, ontology, folksonomy, controlled vocabulary

1. INTRODUCTION

In the context of knowledge work, it is expected that as the available information grows, one would be more effective in his tasks. Unfortunately this is not always the case, and the value of information frequently decreases as the quantity of information increases. This apparent contradiction is due to our human limitations in processing high quantities of raw information.

This paper looks into six patterns for classifying and improving the access to information. Some of these solutions have been used since the 4^{th} century [Wellisch 1994], and are nowadays very well known in the domain of information science. Others came into being on the context of the Web, even though they conceptually share a lot with "older" solutions, but all are used as means for Information Seeking and Retrieval. In one way or another, they can all nowadays be seen pervasively in software systems.

2. AUDIENCE

The main audience for this paper are those wanting to make information quickly reachable, in the context of software systems. Depending on the kind of system, they can be either developers or users of the system. Although the patterns don't lead to a specific implementation, their implications easily crosscut the design of a system, from how the data modeling is done, to how information is perceived and interacted with through the user-interface.

To a lesser extent, we believe this paper may also be useful to those wanting to take their first steps into information indexing, and need to gain a better understanding of the different concepts involved.

This work was supported by Fundação para a Ciência e Tecnologia and by ParadigmaXis, through the grant number SFRH/BDE/33883/2009.
Authors' addresses: F. F. Correia, Rua Dr. Roberto Frias, 4200-465 Porto, Portugal; email: filipe.correia@fe.up.pt; A. Aguiar, Rua Dr. Roberto Frias, 4200-465 Porto, Portugal; email: ademar.aguiar@fe.up.pt

3. ORGANIZING, CLASSIFYING, INDEXING

These three concepts are used throughout the paper, but the differences between them can sometimes be subtle. They can work together to support the same overall goal — to ease the understanding and access to contents. To **organize** is to provide an *order*, that is, to systematize the way in which the contents are recorded and conveyed, so that they can be more easily understood. On the other hand, to **classify** is to assign the contents to *classes*, that is, to group them according to common features — it implies abstraction, and a specific kind of organization. At last, to **index** is to provide the key topics or the classes of the contents as access points to those contents; the emphasis is on how readers can use those common features to actually find and delve into the contents.

The patterns in this paper address these three concerns to some degree. They are Information *Classification* Patterns because they focus mainly on how the different topics of the contents are abstracted and represented.

4. ABOUT THE PATTERNS

These patterns were mined from the experience gathered by the authors while developing software systems — some of them in the information science domain — that use these techniques to make information accessible.

Two approaches to accessing contents — searching and browsing — have proven useful in different contexts. While search provides immediate results, browsing allows an exploratory approach to finding contents, that is key when information needs are ill-defined. The patterns described below focus mainly on supporting the access to contents through browsing.

The first pattern in this paper is the INDEX. Indexes can be elaborate structures, but, in their simplest form, they are lists of terms, usually organized alphabetically. In the context of publishing, the word *index* specifically denotes an alphabetically ordered INDEX of subjects, usually appearing in the back of the document, but unless noted otherwise, in this paper the term *index* is always used in the most general sense, as will become clear in the description of the pattern.

The creation of an index requires the use of a representation language, which supports expressing its entries. When used to represent the information of an index, these languages are called *indexing languages*. The most expressive ones are used for other purposes too, as they are able of representing knowledge in general. Four of the patterns — TAXONOMY, THESAURUS, ONTOLOGY and FOLKSONOMY — are about such languages. Directly or indirectly, they support the creation of INDEXES, and they all strike different balances between the effort of creation, effectiveness of knowledge capture, and ease of use. At last, the CONTROLLED VOCABULARY pattern describes a general approach to disambiguating the meaning of terms; it is key in THESAURI and often used with ONTOLOGIES.

Figure 1 depicts the relations that were just described, and will be explored in greater detail in the description of each pattern.

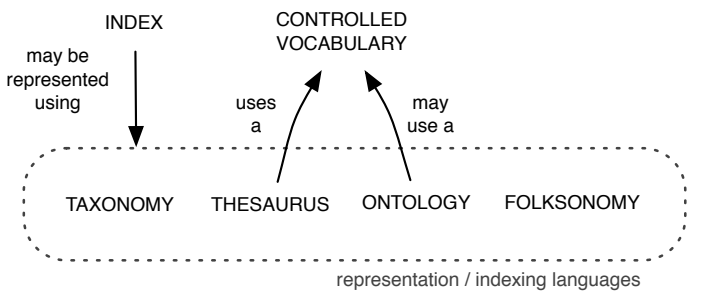


Fig. 1. Information Classification Patterns Map.

- INDEX — Supports readers in finding the contents they seek more efficiently;
- TAXONOMY — Allows representation of information along an hierarchical structure with loosely defined semantics;
- THESAURUS — Provides more semantics and expressivity when representing information, allowing related subjects to be connect;
- ONTOLOGY — Provides even richer semantics and expressivity, and allows representing information as a graph of related subjects, connected through arbitrary types of relations;
- FOLKSONOMY — Supports a collaborative approach to classifying information, but not without a loss in semantics and expressiveness.
- CONTROLLED VOCABULARY – Allows referring precisely to subjects, by disambiguating the meaning of terms.

How the INDEX is represented determines how readers will be able to use it. Its *coordination* is one of the factors to consider during the index creation. During this process, several concepts may be combined to create the index entries, in which case it is called *Pre-coordinate Indexing*. With the opposite approach, *Post-coordinate Indexing*, index entries are based on elemental concepts, which are only pulled together during the access phase, by the reader [Tennis 2008].

Post-coordinated indexes are better to explore the several dimensions of the contents, because they offer no restriction as to which terms to combine, but they don't hint the reader as to which combination of terms might be more interesting to explore. It's also worth noting that, when indexes need to be represented in a static form (e.g., printed on paper), the advantages of Post-coordination are lost, as there is no easy way to obtain the set of contents indexed by more than one term. Pre-coordinate indexes are usually used in such cases [Lancaster and Lancaster 1998, p. 50].

5. GENERAL FORCES

The patterns of this paper share a set of forces, which are depicted by Figure 2. These general forces influence the use of indexing languages and the creation of indexes. They affect almost all of the patterns, and they do so in different ways, as the application context and goal of each pattern also varies. Each of the patterns will describe in more detail the aspects of these forces that matter most in each case.

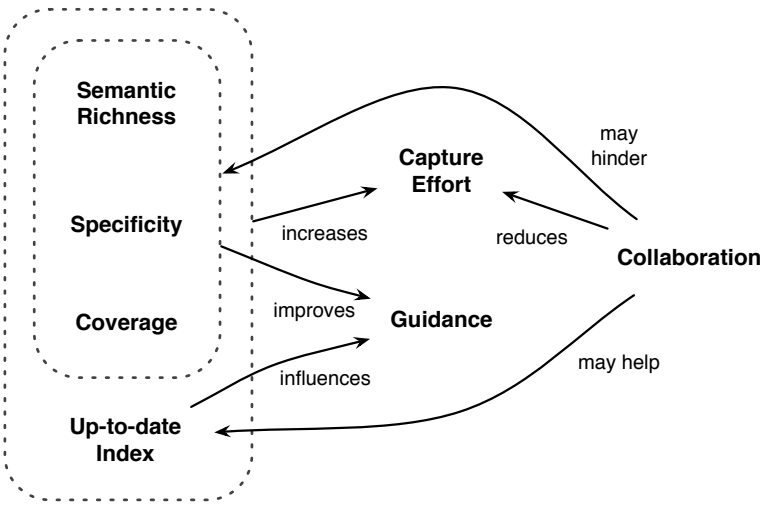
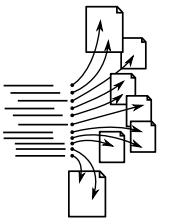


Fig. 2. Relations between the common forces of the patterns

- **Semantic Richness.** How much information does the index contain? Or, more specifically, how unambiguous and meaningful are the index entries? A semantically rich index will be better equipped to guide the reader, although it requires more effort to create. This is especially true if it's created by a group of individuals, as it requires agreement to be reached about the meaning that each entry conveys.
- **Specificity.** Index entries that reflect the same level of specificity of the contents will better align the needs of readers to the contents that they need to be guided to. Namely, entries that are more general should lead to general contents, and entries that are more specific should lead to specific contents.
- **Coverage.** The extent to which the index entries cover the entire contents. By ensuring that the most representative subjects of the contents are part of the index, the readers will be better guided, helping them not to miss useful bits of information.
- **Up-to-date Index.** An index that reflects the current contents makes them easier to find. On the other hand, the reader may never get proficient using the index if it is constantly being completed and updated. Keeping the index of a large and evolving body of information up to date requires a lot of effort, and is usually impracticable unless it's done collaboratively.
- **Capture Effort.** Creating an index may be a costly process. In part, that is due to the time that it takes, but it is also because capturing that knowledge, representing it with elaborate index structures, requires uncommon analysis and abstraction skills. Making it a collaborative process may reduce this effort.
- **Guidance.** An index should be useful for multiple audiences, with different degrees of knowledge on the domain at hand. Readers that are not knowledgeable on a area, and have fuzzier needs, should still be able to use the INDEX effectively. Some knowledge is always needed to carry out the educated guesswork that allows getting to the *right* index entry rapidly. Just a little knowledge of the domain may be enough for a reader to recognize the index entry he needs when confronted with it, although he has to go through the whole index to search for the subject, which can be very time consuming, or even impracticable. But outsiders to the area may not be able to even recognize the entry that would lead them to the desired contents. Building quality indexes, with more and better contextual information around them (e.g., more Semantic Richness, Specificity, Coverage, and ensuring they are Up-to-date) provides more guidance to newcomers to the area, even though it requires more effort from their creators and maintainers.
- **Collaboration.** Obtaining the collaboration of authors and readers of a body of information can be an effective and inexpensive way to create its index and keep it up to date. However, this may happen in expense of quality, as not all participants will possess the skills or be willing to devote a lot of time to the task.

6. PATTERN INDEX



Consider that there is a body of information, which you would like to make available. Due to its extent, one can't just scan it quickly to find what she is looking for. Some topics may appear in multiple places in the contents. And also, after you find a particular piece of content, you might need to come back to it again later.

6.1 Example

Suppose you buy a subscription for an online book, about the very newest and exciting programming language that everyone is now using. Imagine for a moment that the author provided only a list of pages, and the book has no table of contents, nor a traditional back-of-the-book index. You have a considerable amount of pages ahead, but you already know something about the language, so you would like to skip the first sections, and go directly to the first hands-on exercise. You would also like to know where to find help on the language's class library, so you can reference to it as you do the exercises.

6.2 Problem

Without some sort of a guide, and given a non-trivial amount of contents, the effort of trying to go through it all in order to find a particular piece of information may be overwhelming.

As readers, we want to **quickly** find the contents we need. We can go through a document exhaustively searching for what is relevant for us, but again we don't want to take too much time doing it. We may choose to just skim through the whole document instead, but we might miss something relevant if we do so. Moreover, when going through the body of information we have to **repeatedly** assess if each piece of content is relevant or not.

Both **reading** and **evolving** the contents should be easy to carry out, but the most elaborate indices are usually more difficult to keep up to date.

It's important to lower the barrier to **reading** the contents, although that might sometimes happens in expense of how easily they can be **evolved**. Both tasks should be easy to carry out.

We want the index to have a good **breath** of coverage of the contents (scope) and a high **specificity** of its terms (but no more than the contents convey), as this better aligns the contents to the needs of readers. These goals should be balanced with the high **cost** that creating a quality index entails [Lancaster 2003, p. 28–29].

6.3 Solution

Analyze your information to identify the most important subjects, represent them and organize them systematically, making each of them refer back to the contents that they respectively describe.

In this process, make sure to include all topics that are treated in the contents and that may be of interest to the readers. Represent such topics in the index, using terms with the same level of specificity of the contents [Lancaster 2003, p. 36]. How the subjects are actually represented in the index, and the amount of information that the index contains, should depend on how the readers will want to use it.

Instead of going through the whole contents, the information seeker goes through the index *entries*. Each entry has a *locator*, which refers back to the place(s) where the reader may find that subject in the contents. An index may thus be said to be a simplified view — an abstraction — of the subjects that may matter to the reader. Indexes are crafted to allow readers to search through the subjects without having to deal with the whole contents.

By reducing the search universe to a smaller set of terms, we are improving the search **efficiency**, maintaining an equivalent **effectiveness**, because these terms were the result of an analyses process, which selected only the most important topics.

By making the index vocabulary reflect the contents, information seekers will be guided to contents with the same level of **specificity** they have searched for.

6.4 Example Resolved

A Table of Contents provides an overview of its book, but most importantly, it provides an index for the book's sections. So, it supports the reader in finding a particular section, and reaching it easily. Going back to the example presented in the beginning of this pattern, the reader would use the Table of Contents to find the section for the first hands-on exercise, and follow the given page number reference. A traditional back-of-the-book index, which is usually alphabetical, could also help in finding the description of some of the class library functions used in the exercises.

Both — tables of contents and a back-of-the-book indexes — are good examples of INDEXES. They don't have to be used together like in this example, but are often synergistic.

6.5 Known Uses

A menu bar, in a window-based software system, can be considered an index to some extent. It is an abstraction of the system's functional parts, that organize the way users have access to them.

The TeX typesetting system supports the creation of indices, like table of contents, list of figures, list of tables, etc.. The author annotates the contents, and the system will automatically create a list of entries, together with the locators to the respective contents. Printed books often use one, or both, of the INDEXES mentioned in the example sections above. While a Table of Contents of a book indexes its constituent parts (chapters, sections, etc.), a traditional back-of-the-book index indexes the book's contents by subject.

Wikipedia includes several index pages. The *list of wiki software*[1], for example, is a page listing and linking to pages about wiki software packages.

6.6 Related Patterns

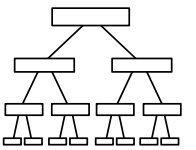
The choice of an INDEX representation language depends on the context at hand. If the contents are not updated very frequently, maybe because they are in print, TAXONOMIES and THESAURI are good options — they assume a closed domain, and that the information of the index is represented before indexing is actually done (i.e., before the creation of the index *locators*). Choosing between the two depends on the complexity of the domain, and on how familiar information seekers are with the topics they will be searching for. The guidance provided by a TAXONOMY may be quicker if the reader already knows the domain area. Newcomers may also prefer a TAXONOMY if the contents are simple enough and the index entries can capture all of the most important pre-coordinations of terms.

The creation of ONTOLOGIES is often motivated by the need to capture and share information with rich semantics, while FOLKSONOMIES, are normally used when the need for collaboration is key. But they are both used to index contents that change frequently. They have appeared and become popular in the context of the Web, possibly because they are more practicable when tools that support such change are easily available.

[1] http://en.wikipedia.org/wiki/List_of_wiki_software

7. PATTERN TAXONOMY

Consider that you want to classify and create an INDEX for a body of information, to support the readers in quickly reaching the contents they need. Readers may have some prior knowledge on the subject they seek, and they can use it to search for the right contents. The frequency at which the contents are updated is not high, when compared with how many times they will be used.



7.1 Example

For some time, Amy has been keeping files in the “Documents” directory of her computer. She has already dozens of documents, a great many of them created by herself. She now needs to find a work related document — a project proposal she sent to Mr. Smith during the past year.

She doesn’t know the exact name of the file, so she tries narrowing her search. She orders files by date, but that still leaves her with too many files to go through. She also tries a text search, but that retrieves all the invoices she kept from her landlord — John Smith — as well as dozens of letters from one of her suppliers at work — Smith & Co..

7.2 Problem

Without knowing the exact term, it can take too much time to search through the whole index.

To sort out the index entries they seek, readers need additional information, which describes and contextualizes the entries, and that they can use to partition them. Although a **semantically richer** index implies a greater **effort** from indexers, it better **guides** the readers in finding contents.

7.3 Solution

Organize the index entries hierarchically. The meaning of the relations between parent and child entries may vary, so the index can be partitioned by different dimensions into several subareas.

Choose entries that cover the whole domain , and try to keep the taxonomy tree balanced. Add entries that can be easily understood by the readers when taken in context with the upper taxonomic levels. When creating the entries of a taxonomy don’t try to make them stand on their own. To assemble meaning from an entry, readers will consider its context in the taxonomy. The same terms may convey different meanings when they appear in different points of the taxonomy.

Each taxonomic level relates to the upper level according to one of its dimensions. In case the same term is placed in more than one point of the taxonomy it does not mean that the same subject is being classified in multiple ways, but rather that different subjects are being represented. The order in which such dimensions are represented as parent-child entries should reflect the knowledge that we foresee readers may have, and the way they will seek the entries in the taxonomy.

From an indexing point of view, a TAXONOMY can be said to be a *fixed-vocabulary* language, as a pre-established representation of terms is taken as a basis for the indexing process. Taxonomic indexes are *pre-coordinated*, because each entry is a combination of terms, that describe a group of other entries.

The partitioning of the index along a tree structure makes navigating it more **efficient**, as readers are able to eliminate from their search several index entries at once, when they belong to a subarea that does not interest them.

Also, the more a reader knows about a given domain, the more **efficient** she is navigating through a taxonomy of that domain — she will be quicker in grasping which dimensions the taxonomy is using to partition the index, and how she should navigate it to reach the intended index entries. Newcomers may have to explore the index first, before being able to use it efficiently.

But taxonomies are not without liabilities, and the level of **expressiveness** that they allow is one of them. In practice, it may be hard to group contents according to a single sequence of dimensions. Although you can try to anticipate which features of each piece of information will be the most relevant to future readers, different readers may easily have very different needs.

The added **effort** of pre-coordinating and grouping together related terms makes a TAXONOMY harder to create and maintain when compared to using a plain list of terms. Only considering a high number of readers, and a low rate of updates does such effort pay itself easily.

7.4 Example Resolved

Realizing that she must stop going through all the files every time she needs one of them, Amy started organizing the files into subdirectories. She created three subdirectories inside the "Documents" directory: "Family", "Work" and "Friends". Inside the "Work" subdirectory she created some more subdirectories, one for each customer.

Whenever she needs to reach the project proposal for Mr. Smith again, she will rely on the directory structure to guide her. She will first open the "Work" subdirectory, then the "Mr. Smith" subdirectory, and finally go through the files there.

The top-most directories partition the larger groups of files, according to Amy's social groups (family, work, friends). In turn, the directories inside the "Work" directory are grouping the files by person — sender or receiver. Each level uses the dimension that better helps Amy navigating that particular group of files.

7.5 Known Uses

Several software systems use the concept of "folders" — container of items, used to organize them. Such items can themselves be other folders, and thus contain other items, forming a tree-like structure, that can be regarded as a TAXONOMY. An example of such use is the Alfresco Content Management System[2], which has the concept of "Spaces". Alfresco Spaces are generic containers, that behave much like "folders".

A Web Directory, like the Open Directory Project[3], can be seen as a TAXONOMY that classifies websites on the World Wide Web. Web Directories were once important to find Web resources, but due to their very high rate of change and growth, very few have survived in favor of full text search engines.

A Table of Contents, either digital or in print, may be seen as a TAXONOMY that organizes contents according to sections and chapters.

The Dewey Decimal Classification System is one of the TAXONOMIES with the most widespread use. Its goal is to cover all areas of knowledge, supporting the classification of books and other library items, and providing a way to easily find them, on the online catalog or shelves of a library.

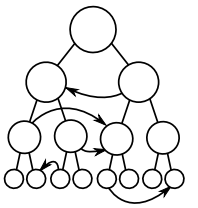
7.6 Related Patterns

TAXONOMIES, like THESAURI, ONTOLOGIES and FOLKSONOMIES, can be used to represent an INDEX. Like THESAURI, TAXONOMIES assume a closed domain.

When information seekers are not newcomers to the domain area, or if the contents are simple enough, TAXONOMIES allow to reach information quicker, but otherwise, semantically richer indexing languages will provide better guidance.

[2] http://www.alfresco.com/
[3] http://www.dmoz.org/

8. PATTERN THESAURUS



Consider that you want to classify and create an INDEX for a body of information, to support the readers in quickly reaching the contents they need. Readers may have **very different backgrounds and different levels of prior knowledge** on the subject they seek. The frequency at which the contents are updated is not high, when compared with how many times they will be used.

8.1 Example

Suppose a library keeps a set of documents about art, and allows readers to find them through an online catalog. John would like to find some documents about a famous painter, whose name he doesn't recall right now. All he remembers is that the painter was contemporary to Monet, and painted using the same style. Paul happens to also be looking for documents about the same painter, but all he knows is that he created the painting *"Dance at Le Moulin de la Galette"*.

Both John and Paul need to do some research before getting to the documents they need using the library's index. John will start by finding documents about Monet. He then uses these documents to learn that Monet was an impressionist, and then find documents about the painters of that movement, to finally recognize Renoir as the name he was missing. Searching by Renoir on the library's online catalog will finally reveal the documents he needs. Paul, on the other hand, will first need to search for the name of the painting he knows, to find that it was painted by Renoir. He can then use the library's online catalog effectively.

8.2 Problem

Different people seeking the same contents need the index to guide them in different dimensions.

The same contents may need to be accessed differently, depending on the knowledge of the reader. To sort out the index entries they seek, readers need information that describes and contextualizes those entries. Although a **semantically richer** index implies a greater **effort** from indexers, it better guides the readers in finding contents.

8.3 Solution

Organize the index entries as a network of subjects. Define the meaning of each subject carefully, by using a CONTROLLED VOCABULARY**, and connect them with other, related, subjects.** More specifically, organize subjects according to five different elements/connections:

- Broader/Narrower — Thesauri, like taxonomies, are organized hierarchically, but the semantics of such relations is better established than with taxonomies. *Parent* subjects are said to be *broader*, and child subjects *narrower*, in the sense that the scope of each child subject is narrower than the scope of the parent.
- Scope Note — Each subject represents not merely a term, but a concept, that is part of a CONTROLLED VOCABULARY. The meaning of the concept is defined through a *scope note*.
- Synonyms — Other terms that may describe the same subject. Synonyms are *unauthorized forms* of the CONTROLLED VOCABULARY used to support the THESAURUS.
- Topmost — Each subject has at least one topmost subject, which is the one that would be found by following the *broader* relations until the broadest possible subjects are reached.
- Related — Refers to related subjects, that are not broader or narrower.

Despite THESAURUS' entries representing concepts, the terms are usually emphasized more than the underlying concepts, and THESAURUS are very often perceived as just a set of connected words. A THESAURUS allows a richer description of subjects when compared with a TAXONOMY, as it supports expressing broader/narrower relations, which have more concrete semantics than the hierarchical relations of a TAXONOMY. *Related* connections

support expressing other (unspecific) kinds of relations. In spite of the added expressiveness when compared with TAXONOMIES, THESAURI are sometimes extended with even further attributes and kinds of relations.

To create a THESAURUS, identify subjects (i.e., index entries) at different degrees of abstraction, reflecting the different levels that may be found in the contents, from the very coarse-grained (general) subjects to the very fine-grained (specific) ones.

Represent in the THESAURUS all the knowledge in the field, and reuse is as often as needed. You should index the same piece of contents with multiple index entries, providing multiple access points, possibly to be combined, when searching the index.

THESAURI provide semantically richer connections between subjects, which makes them **quicker** to navigate for those without much prior knowledge on the domain.

THESAURI are meant to comprehensively cover their target domain, and are reviewed and updated only sporadically. They are, for this last reason, called a *fixed vocabulary*. This supports their **usability**, as it makes it easier for readers to learn how the thesauri that they use are organized.

Expressiveness is better than with a TAXONOMY, but THESAURI demand more attention to semantics, which can imply more **effort** during creation and maintenance. However, in practice, maintenance may not actually be harder than with a TAXONOMY, because the meaning of each THESAURUS' entry is more fine grained and better defined — it's easier to improve an entry while being confident that the rest of the THESAURUS remains consistent.

8.4 Example Resolved

Suppose that the catalog software, of the the library of the example above, allows thesaurus-based indexing. The librarians have decided to built a thesaurus in the art domain, and are using it to index the documents that they are curating.

John will start looking for the elements he knows. He will first seek for the term "monet". He finds the corresponding subject in the thesaurus that he confirms to be the one he is looking for, upon reading the scope note, and by observing that it is *narrower* term of the "Painter" entry. He quickly goes through that thesaurus entry, and finds out that "impressionism" is a *related* thesaurus entry. He then looks at the remaining entries related with "impressionism", and recognizes "Renoir"as the painter he was seeking. He now just needs to follow the index locator(s) for that entry, to reach the documents about Renoir.

Paul, on the other hand, will start seeking for the term "Dance at Le Moulin de la Galette". He finds it, and too sees "Renoir" as a related entry. All he has to do now is follow the locator(s) to reach the documents he needs.

8.5 Known Uses

GISA is a software product for creating records and descriptions of archival documents, which uses a thesaurus-based index. Its Web frontend can be found in the websites of several Portuguese archives, like the Archives of the City Hall of Gaia[4] and the Archives of the University of Porto[5], among others.

The *Index New Zealand Thesaurus*[6] was created to describe publications about New Zealand and the South Pacific in the areas of social sciences and humanities. It provides access to journal and newspaper articles.

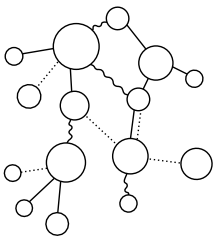
8.6 Related Patterns

THESAURI can be used to represent an INDEX, like TAXONOMIES, ONTOLOGIES and FOLKSONOMIES. Like TAXONOMIES, THESAURI assume a closed domain.

THESAURI are better at guiding information seekers than TAXONOMIES. The entries of a THESAURUS form a CONTROLLED VOCABULARY in the sense that their meaning is established unambiguously.

[4] http://arquivo.cm-gaia.pt/
[5] http://gisa.up.pt/pesquisa/
[6] http://innz.natlib.govt.nz/content/thesaurus/

9. PATTERN ONTOLOGY



Consider that you want to classify and create an INDEX for a body of information to support the readers in quickly reaching the contents they need. Readers may have very different backgrounds and different levels of prior knowledge on the subject they seek. **Contents may frequently be created and updated, and they are likely accessed through platforms that enable collaboration, such as the Web.**

9.1 Example

A research institute has, over time, produced a large body of information. Some of these contents were recorded by the research groups themselves, in different software systems that they maintain. These contents are organized in different ways, depending on the system they were captured in; they have varying levels of structure, from information systems, to text documents, to raw experimental data; and they keep evolving as more results are found and documented.

Linda is researching on the area of automated software testing, and would like to know which results her colleagues have achieved in this area in the last few months. She hopes to find work to build upon, or find researchers of other groups to collaborate with.

The institute provides a list of systems that it uses to capture contents internally. Each group works on a specific subarea, and knowing this could help Linda find the systems with the contents she needs. However, there aren't groups working specifically on automated testing, and almost all of the groups have done some automated testing at some point. If she wants to make sure to find all the contents she needs, Linda will have to search through all the systems and their information.

9.2 Problem

Without a rich and accurate representation of the contents, an index is not able to guide a reader effectively.

To sort out the index entries they seek, readers need additional information, which describes and contextualizes such entries. Although a **semantically richer** index implies a greater **effort** from indexers, it better **guides** the readers in finding contents.

9.3 Solution

Organize the index entries as a network of subjects. Define the meaning of each subject, and connect it with other, related, subjects. More specifically, organize subjects according to elements such as *individuals*, *classes*, *attributes* and *relations*, among others.

An ONTOLOGY is a *formal, explicit specification of a shared conceptualization* [Gruber 1993]. It may gather a collective understanding on a given area and be open to, and constantly updated by, a group of people. Several ontology languages exist, but common components include *Classes*, *Attributes*, *Relationships* and *Individuals*. Subjects may be defined by classes or individuals, and are characterized by attributes and relationships. Attributes and Relationships are themselves described by classes, and this mechanism allows the language to be extended as needed. Given its formal nature, an ONTOLOGY is very fit to use a CONTROLLED VOCABULARY.

The ability of expressing virtually any kind of attribute and relation makes ONTOLOGIES able to provide a **richer description** of subjects when compared with a TAXONOMY a THESAURUS or a FOLKSONOMY, and has thus the capacity of **guiding** information seekers more effectively.

Even though an ONTOLOGY may be **open** and constantly updated by a community, it may sometimes prove to be a difficult endeavor to reach a **consensus** over the conceptualizations.

9.4 Example Resolved

Going back to the example presented in the beginning of this pattern, the institute can build an aggregator, that gathers contents from each system, and provides a unified and abstracted representation of them. This ontology can be used as a global index, that users use to reach the actual contents. Although the source contents need to be semantically rich to be aggregated, the ontology can be completed with additional information to make other contents accessible through the index too.

Linda now uses the ontology-based index to find an entry about automated testing, and follows the index locators to documents maintained in a system used by the Artificial Intelligence Group, and to some unit-test coverage data that was used for creating software visualizations by the Computer Graphics Group. Linda is directed to the right systems and, more specifically, to the right contents within the system.

9.5 Known Uses

Semantic MediaWiki[7] is an extension to the MediaWiki wiki engine. It adds the ability to annotate the contents of a page, conferring it semantics. This data can then be queried, by one or several of its dimensions, and the results provided, from within a wiki page, as an access to the contents in question. Among other features, it supports exporting data as OWL[8], an ONTOLOGY representation language based on XML.

In research, we can find several approaches to indexing contents with an ONTOLOGY. An example, among several others, is the work by Luances et al, which uses an Ontology to improve the query capabilities to a Geographic Information System [Luaces et al. 2008].

Plone Ontology[9] is an add-on for the Plone Content Management System, that allows to collaboratively create an ontology that can be navigated and used to access the system's contents.

9.6 Related Patterns

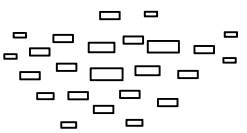
ONTOLOGIES can be used to represent an INDEX, like TAXONOMIES, THESAURI and FOLKSONOMIES. Like FOLKSONOMIES, ONTOLOGIES assume an open domain.

If a rich description of contents is more important than supporting collaboration, an ONTOLOGY makes a better indexing language than a FOLKSONOMY.

The meaning of the elements of an ONTOLOGY is established unambiguously, and in that sense, it may be very close to using a CONTROLLED VOCABULARY. However, that's often not the case, as the elements of an ONTOLOGY don't necessarily have an authorized form.

[7] http://semantic-mediawiki.org/
[8] http://www.w3.org/TR/owl-features/
[9] http://plone.org/products/ploneontology

10. PATTERN FOLKSONOMY



Consider that you want to classify and create an INDEX for a body of information, to support the readers in quickly reaching the contents they need. Readers may have very different backgrounds and different levels of prior knowledge on the subject they seek. **The amount of contents is overwhelming;** they may be frequently created and updated, and are likely accessed through platforms that enable collaboration, such as the Web.

10.1 Example

Suppose you have created a software platform for amateur photographers through which they can publish their best works on the Web. You want to let users easily seek the contents they need, but you don't have the resources to hire a team to manually classify such a large set of pictures, and you also don't want to demand from users a lot of effort to classify their photos by topic.

Kate, an early adopter of the system, has just uploaded her photo album from the last five years, and she is now wondering how to find the pictures of the weekend she spent last year in Portugal. She would also enjoy knowing what other pictures of Portugal there are on the system, as she would like to find new places for her next visit.

10.2 Problem

Indexing and classifying a large body of information is unfeasible or at least very costly to carry out.

Assuming the information is already recorded, one could consider assigning a team with the task of creating an index to ease all subsequent accesses. However, the **effort** of such an endeavor is usually very high. This is aggravated if such information is in constant **change**, in which case, the classification efforts cannot be limited in time, and have to follow the entire lifecycle of the information.

Those that are most **knowledgeable** about some specific contents are not external indexers but its own creators, as they are more aware of its context and domain. On the other hand, information creators are not necessarily aware of what makes a good index, and may lack the necessary **analysis and abstraction skills** to represent elaborate index structures.

10.3 Solution

Ask the creators or users of the information to identify the set of words that most accurately describe the contents, and *tag* them with those terms.

Those words are the *entries* of the index. By letting — and encouraging — users to assign descriptive terms to pieces of information, an index will emerge. It will not be defined up-front, but rather will gradually appear from the practice of collaboratively tagging contents [Furner 2010].

There is not a single way to seek contents using a folksonomy. Tags can be made available to information seekers as simple alphabetically ordered lists, or as *tag clouds*. Tag clouds present the several tags by laying them out in different locations, and using different font sizes and colors to highlight the relative importance of each one, usually directly reflecting the number of times they were used to tag some content (i.e., the number of underlying *locators*).

The final result is a *word index*, as opposed to a *subject index*. This kind of index does not make use of a CONTROLLED VOCABULARY, and thus its entries lack a **strong semantics**, leaving to the reader the job of figuring out if the contents tagged with a given entry actually refer to what he is looking for. However, this is also one of the biggest strengths of this solution; by reducing to a minimum the **effort** required in the analysis phase, the creation of the index is easy enough to be done by any information creator, in a **distributed** way.

10.4 Example Resolved

Going back to the example presented in the beginning of this pattern, the developers have just added to the system a feature that allows users to tag their own contents. Kate can now tag her photos with any term that she finds descriptive.

She adds the “portugal” tag to the photos that she took in Portugal. It happens that other users are using that tag too, so it doesn’t take too long for Kate to be able to find other photos of this country. She will just follow the link on the “portugal” tag of one of her photos, and be lead to a full list of photos tagged with this term.

10.5 Known Uses

Folksonomies are very popular on the Web, and used by several successful websites.

Flickr[10] is an image and video hosting service in which users can classify their photos with tags. The community can then search photos by their tags, or they can seek photos through a tagcloud. Delicious[11] is another well-known service on the Web that makes extensive use of tags as a way to bookmark and describe Web pages.

Despite often supporting both mechanisms, many weblog engines favor the use of tags instead of categories to classify posts. Wordpress[12] is a blog engine that supports both approaches, and can provide access to its posts and pages through tagclouds.

Some systems were not designed to use tags, but are extended with plugins that add this capability. An example is the Trac software-forge[13], and its Tags plugin[14]. Trac integrates several features useful for software development, like a wiki engine, source-code browser and issue-tracking system. The Tags plugin allows to label wiki pages, which can then be easily queried to create indexes.

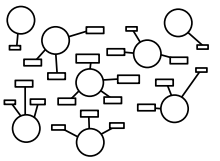
10.6 Related Patterns

FOLKSONOMIES, like TAXONOMIES, THESAURI and ONTOLOGIES can be used to represent an INDEX. Like ONTOLOGIES, FOLKSONOMIES assume an open domain.

If supporting collaboration is more important than a rich description of contents, a FOLKSONOMY makes a better indexing language than an ONTOLOGY.

[10] http://www.flickr.com/
[11] http://www.delicious.com/
[12] http://wordpress.org/
[13] http://trac.edgewall.org/
[14] http://trac-hacks.org/wiki/TagsPlugin

11. PATTERN CONTROLLED VOCABULARY



Consider that you want to classify and create an INDEX for a body of information, to support the readers in quickly reaching the contents they need. The representation of the index, and of the information itself, is done using text. This means it can take on several meanings, but when readers seek the content of a subject, they have a specific meaning in mind.

11.1 Example

George is using an electronic address book, where he keeps all of his contacts, including customers, suppliers and friends. The address book allows several kinds of contacts (telephone, email, etc.), but he is using it to store mostly phone numbers.

Today George decided to call his friend Richard Smith, and he was faced with an issue: there are two Richard Smiths in the address book. One of them is surely his friend, and he thinks the other might refer to one of his suppliers, that has the same name. There's also a contact with the name "Ringo" on the list. It's an alternative contact of George's friend, but it's under a name that he didn't remember to check.

11.2 Problem

A single term or expression doesn't convey a precise concept, and can be interpreted to different meanings.

When going through a list of terms we read them in the context of the that list, and the way it is being presented to us. A textual expression that is not **semantically rich**, may make it difficult for readers to assemble the same meaning that was originally intended. Clearly establishing the meaning of each term will imply a higher **effort** to create and maintain that vocabulary.

Furthermore, two different terms may be intended to convey the same meaning. Such **ambiguity** may make information consumers unsure if two apparently different terms actually mean the same or different things. On the other hand, by representing the same concept with different terms, we are supporting **different information consumers**, which may be looking for the same meaning using different terms.

11.3 Solution

Select an *oficial* term to denote each concept, and use it every time you need to refer to that concept.

Such term is called the *authorized form* of the concept. Pick the most descriptive term (or expression) for the authorized form, making sure that no two concepts share the same term. If the same term is generally used for different concepts, explicitly add a qualifier to the term, to resolve the ambiguity.

Also include aliases of the term in the controlled vocabulary, as *unauthorized forms*, and use them to find their corresponding authorized forms when necessary. When listing the terms of a controlled vocabulary, emphasize the authorized forms, so that the users of the controlled vocabulary can favor them when referring to concepts of the vocabulary. You can show unauthorized forms only on request, and/or format them differently.

Unlike the terms of natural language vocabularies, a CONTROLLED VOCABULARY consists of a list of terms preselected by the author of the vocabulary, which relates the different words that represent the same thing (synonyms) and distinguishes the different concepts with the same name (homographs and polysemes) [ANSI/NISO 2005].

A CONTROLLED VOCABULARY reduces language **ambiguity** by establishing a single authorized form for each concept. Controlling a vocabulary requires some **effort**, and some **collaboration** between those envolved in its creation, to achieve a shared understanding of the concepts underlying the authorized forms.

11.4 Example Resolved

George's address book supports having several contacts assigned to a name, and in order to avoid facing the same problem in the future, George takes some time to reorganize it.

He confirm a few numbers, and starts merging records that refer to the same person. He creates an entry with the name (i.e., authorized form) "Richard Smith", to which he associates his friend's work phone number, and personal phone number. The address book also allows to add aliases (i.e., unauthorized forms), and he adds "Ringo" as an alias for that entry, should he search for that term in the future.

To create an entry for his supplier with the same name, he needs to disambiguate the authorized form using a qualifier, and uses "Richard Smith (suplier)" as the authorized form for that contact.

11.5 Known Uses

When using an information system, users often need to fill in fields from a predefined list of possible values, sometimes presented with a combo-box graphical control. The terms can often be edited only by the administrators of the system, which should choose them carefully to avoid misinterpretation by the users. In this sense, these lists of preselected terms can be seen as controlled vocabularies, even though usually only one form is supported (the authorized form), and it's not possible to express aliases for each term (i.e., add unauthorized forms).

The PSH[15] — Polythematic Structured Subject Heading System — is a controlled vocabulary and an indexing system, built by the Czech Republic's National Technical Library. It was created for indexing and searching contents by subject.

11.6 Related Patterns

By eliminating ambiguity and controlling synonyms, CONTROLLED VOCABULARIES support the creation of INDEXES, providing information consumers a better guidance. Some indexing languages, like THESAURUS and ONTOLOGY, imply a strong definition of the concepts behind the terms, and are naturally a good fit for the use of a controlled vocabulary. Others use a different approach — TAXONOMIES and FOLKSONOMIES rely on an implicit definition of the underlying concepts through the context in which the terms appear.

[15] http://www.techlib.cz/cs/564-english-version/

12. CONCLUSION

In spite of the vast amount of available information around the topics described in this paper, some of these notions are not consensual; especially those of TAXONOMY, THESAURUS and CONTROLLED VOCABULARY. The patterns here presented focus on what the authors believe is the essence of these concepts, but it's uncommon to find a "pure" TAXONOMY or THESAURUS, as they are very often extended with other descriptive elements, that provide the expressiveness that indexers need in their specific context. Moreover, an ONTOLOGY can be seen as an extended (i.e., more expressive) THESAURUS, and a THESAURUS can be seen as an extended TAXONOMY, which further blurs the border between them.

The context of use of these concepts has changed throughout time, and this has helped to dilute their meanings. Before computers were part of our lives, knowledge representation was mostly subject to what paper would allow us to represent, and how access to that information could be provided. Nowadays, *representing* and *accessing* information aren't so strongly dependent on each other. With computational support, authors can collaborate to create semantically rich indexes, without hindering the access to contents.

Traditional indexes assume a closed domain, which makes them suitable for static information. However, nowadays, the scope of some bodies of information is much more difficult to outline, and may easily keep growing and morphing [Wright 2005].

The existence of large collections of contents is increasingly more common, making it more likely for concessions to be made in terms of semantic richness, in favor of a better coverage and a collaborative approach to indexing [Voss 2007]. The accessibility to contents is ever more dependent on the care and efforts of their creators, which means that semantically richer indexes may only be practicable with indexing tools that support the sense-making and agreement process between the several stakeholders.

13. ACKNOWLEDGMENTS

The authors would like to thank Joe Yoder for his support and valuable comments while shepherding this paper for PLoP 2011, and to Fernanda Ribeiro, who provided interesting insights in the domain of information science. We would also like to thank both Fundação para a Ciência e Tecnologia and ParadigmaXis, which partially financed this work through the grant number SFRH/BDE/33883/2009.

REFERENCES

ANSI/NISO. 2005. Z39.19. Guidelines for the construction, format, and management of monolingual controlled vocabularies.

FURNER, J. 2010. Folksonomies. In *Encyclopedia of Library and Information Sciences* 3rd Ed., M. J. Bates and M. N. Maack, Eds. CRC Press, Boca Raton, FL, USA, 1858–1866.

GRUBER, T. R. 1993. A translation approach to portable ontology specifications. *Knowledge Acquisition 5,* 2, 199–220.

LANCASTER, F. W. 2003. *Indexing and Abstracting in Theory and Practice* 3rd Ed. University of Illinois Graduate School of Library and Information Science.

LANCASTER, F. W. AND LANCASTER, F. 1998. *Indexing and Abstracting in Theory and Practice* 2nd Rev Ed. Graduate School of the University of Illinois.

LUACES, M., PARAMÁ, J., PEDREIRA, O., AND SECO, D. 2008. An ontology-based index to retrieve documents with geographic information. In *Scientific and Statistical Database Management*. 384–400.

TENNIS, T. 2008. Organization of information and resources — lecture materials on pre- and post-coordinate indexing.

VOSS, J. 2007. Tagging, folksonomy & Co-Renaissance of manual indexing?

WELLISCH, H. 1994. Indexing. In *Encyclopedia of Library History*. Garland, New York, NY, USA, 268–270.

WRIGHT, J. 2005. The future of indexing? http://www.writersua.com/articles/indexing_future/.