УДК 002:025.4

INFORMATION ORGANISATION IN SUBJECT GATEWAYS

Jadwiga WOŹNIAK-KASPEREK

Institute of Scientific Information and Book Studies of Warsaw University,
69, Nowy Świat Str., Warszawa, 00-046, Poland
e-mail: jbwozniak@uw.edu.pl

The article is dedicated to one of the key problems of modern society, the optimization of informational search. The author dwells upon the necessity of experts taking part in the selection and indexation of information. Two additional factors in solving the problem of informational search are academic dictionaries and uncontrolled sets of key words.

Key words: informational search, academic dictionaries, knowledge organization systems, subject gateways.

Inside each one of us there dreams, though we sometimes fail to realize it, an instinct for understanding: we are eager to comprehend, explain, prove. But noting can be grasped, clarified, or demonstrated without appropriate prior knowledge. Still, nowadays obtaining information (unfortunately not knowledge) is peanuts. What is lacking in our memory we can easily find on the Internet; sometimes at the price of financial cost. But, is it really as simple as that, or are we in for disillusionment? The average Internet user is at risk of accepting the false impression that search engines, such as for instance Google, can find absolutely everything. Indeed search engines find a lot; more than that: they find more and more every day, although not everything they come across is of the same value, not to mention the mass of completely irrelevant documents they bring forward. On the one hand, the user is drowning in the unlimited expanse of answers, but on the other, if he is interested in a specialised problem the search engine frequently comes up with no answers at all! Does that mean that the Internet has no information on the given problem? Sometimes it's true, but more often it is not, the information is there, while for numerous reasons we are incapable of finding it.

For this reason optimisation of information retrieval has become one of the key information problems of contemporary times. Information science is also hard at work wrestling with this issue, conducting theoretical research and practical experiments, both with respect to retrieval methodology, organisation of information and indexing. The way this matter is settled will resolve in the future (or indeed is already deciding today!), whether the vast quantities of information stored on the Internet will be easily accessible, or will remain only a potential, which is already classified as the "deep", or "invisible" Internet.

The majority of this "invisible" content is comprised of valuable, specialised databases, archives of important source materials, bibliographies, full-text reviews of scientific journals, archive versions of press articles, scientific abstracts, archives of discussion groups, professional thesauri and encyclopaedias, pictures, audio and video files, graphic files, software, telephone/address directories, etc. These databases are created and kept among others by higher

educational institutions, libraries, and research institutes of national or international character. The "invisible" content is verified, evaluated and prepared by editors*, who in most cases are highly qualified specialists in their fields. Some cover numerous scientific disciplines, while other, are specialized subject gateways registering and organising collections of information relative to a particular theme. Subject gateways often are the only one way, which enables access to information far more abundant than normally available. Information resources available via subject gateways are professionally selected and organised in accord with the rules of the indexing language and respecting intellectual rights. To avoid any misunderstanding let me repeat it once again: information in these services comes from the Internet, it is selected by experts, and by them or with their participation is indexed and organised. Subject gateways also provide evidence that information in the Internet is not indexed and researched with the sole aid of free key words. This is due to the fact that in many cases (if not in the majority) control is attained by means of indexing systems or the so called knowledge organisation systems.

Let me now discuss in more detail the organisation of information in subject gateways. I would like to demonstrate that the best librarian and information science practice can join forces with the potential of contemporary information technology. I will confine myself to presenting models. I will not be providing here any instances of selected services. I am perfectly sure that the reader can easily find in the Internet numerous exemplifications of the model solutions I demonstrate[1]. The article is aimed at demonstrating (with respect to subject gateways) that the two sets of tools for representing and retrieving information, i.e. free key words and controlled vocabularies are complementary; this complementariness, in my opinion, is not the optimal solution, but it is far more effective than complete lack of any control.

Free keywords

Hasty observations of Internet users' information behaviour result in the growing number of opinions, which proclaim the approaching replacement of all types of vocabulary control by uncontrolled keywords. Doubtless, keywords are universally used in web searches, but that does not necessarily imply the lack of any kind of dictionary, even in the form of a synonym's ring. If in response to a typing error the system asks: "Do you mean...", the question results from the fact that the engine possesses some kind of vocabulary knowledge (the problem, how this dictionary knowledge is expressed, is of course a different matter), thanks to which it knows that the semantic value of e.g. "cataloging" is equivalent to "cataloguing". Rosenfeld and Morville devoted a lot of attention in their *Architektura informacji*[2] to thesauri, controlled vocabularies and metadata, demonstrating how a controlled dictionary enables combining various systems (organisation, labelling, navigation, searching, etc.) of a single service, in order to enhance its functionality and effectiveness. In my opinion, reasonable control of vocabulary is both necessary and desirable. I think that opinions which claim that key words are a universal cure for all searching problems are dangerous. Keywords are important, necessary and indispensable, but are not the only and exclusive tool available

* Decency requires emphasizing that in the case of services constructed "manually" by teams of editors, a limited contents of such a database and relatively high costs must be considered.

[1] To begin with, let us consider the list provided by Lidia Derfert-Wolf: Derfert-Wolf L. Serwisy tematyczne o kontrolowanej jakości w Internecie – subject gateways (Wykaz: *Biuletyn EBIB* [on-line]. – 2004. – Nr 6 (57). – http://ebib.oss.wroc.pl/2004/57/wykaz.php [Aug. 12, 2009]).

[2] Rosenfeld L., Morville P. *Architektura informacji w serwisach internetowych*. – Gliwice: Helion, 2003.

and advisable in use. They are a complementary solution, an additional proposition for effective use in certain information-retrieval situations, e.g. in uncontrolled full text searches. The user is entitled to the comfort of using keywords in his dialogue with the system, but it is our duty to make sure that behind the scene the transparent knowledge system assists the search processes going on in the foreground. We shouldn't do away with professional experience and results of scientific research. And I am not advocating uncritical copying of established models. WWW services, intranets, and digital libraries function in an environment, which requires new and in a certain sense revolutionary solutions, which at the same time will create conditions favourable for such solutions. It would be a sin not to take advantage of this opportunity. But, it would be no lesser sin to destroy the achievements of subject cataloguing only because they stem from a pre-Google and pre-Amazon era.

Knowledge organisation systems

Division of knowledge organisation systems (KOS) is typically done on the basis of how the ordering of headings which form the semantic field of the discipline is done. Two basic types of KOS are discerned in professional literature: exact organizational schema and fuzzy, ambiguous organizational schema.

Exact schema dwell on the formal criteria of ordering, while ambiguous schema on the logical and semantic ones. Exact schema offer quick and easy information searching, providing the user knows the names of the persons, places, works or other information objects. However they do not offer the possibility of scanning the whole content of the source.

In the case of ambiguous schema we are dealing with division into classes, which are arranged according to semantic and logical criteria. Among such schema one finds various hierarchical structures, the classes of which are represented by expressions taken from a natural language. Because it seldom happens that an expression has only one meaning, in the process of creating a hierarchy only one of the connotations arbitrarily chosen is considered. Another solution is considering all the meanings of the expression.

Among the many well known KOS models the most frequently used are:

– in the category of exact schema – lists of terms;

– in the category of ambiguous schema – classifications and categorisations (hierarchical structures), thesauri and ontologies (relational lists).

Lists of concepts (terms)

A list of synonymous expressions in the form of a ring of synonyms or a list of preferred headings in the form of an authority file is the simplest controlled vocabulary. In library and information activities the most frequently used variants of concept lists are authority files and indexes. If a hierarchical relationship is established among the elements of a list than we are dealing with a classification schema.

Classifications and categorizations

Direct indication of hierarchical relations between headings is the chief characteristic feature of this group of KOS. But KOS, sometimes referred to as classifications, frequently do not fulfil the conditions required of a classification in the logical sense of the term; however this does not negatively influence their usefulness. In practice they are various structures, which exploit parallel divisions according to subjects and disciplines. Information architecture considers subject heading systems as hierarchic structures. Subject heading

systems themselves should be considered as exact lists of controlled terms. But on the level of organising information resources subject headings lead to constructing hierarchical structures*, where the most general classes are represented by simple (not divided) headings, which name the subject of the document (sometimes also its form – so-called form/genre headings), while subordinated classes are discerned by subdivisions, which define aspects. In other words, the subdivisions, after being joined with headings (which results in so-called complex subject heading), generate a certain hierarchical structure of detailed issues, which are subordinated to simple headings.

In current practice we deal ever more often with categorisations[3] (not classifications) as forms of information organisation. Classifications are of a differentiating nature and are set to find the features which distinguish one thing from another, while categorisations are directed at linking, looking for the common, or the similar.

The notion of categorisation is ambiguous. Usually this term applies to: 1) the process of differentiating of objects and their collections treated from a given point of view as homogenous, i.e. the process of arriving at categories; 2) the process of ascribing of objects to already existent categories. Categorisation can also be understood as the synonym of the term "category", or more precisely: system of categories.

Categorisation lends itself freely to research by cognitivists, who link it with the so called classical theory of categorisation** and use it as background for presenting their own, in their opinion, more adequate propositions for solving the issues they deal with. Contemporary linguistics emphasize the specificity of defining the world by individual languages. In other words, each individual language creates an alternative vision of the world, which is specific for this language and represented in its system of semantic fields. Only part of this system belongs to classical Aristotelian categories. The specificity of individual languages is reflected primarily in their own segmentation of the world by wording. Let us consider the Polish notion of "mgła", which has three equivalent semantic classes in the English language: "mist", "fog" and "haze". On the other hand, the English term "cloud" finds two Polish words: "chmura" and "obłok". The Polish word "śnieg" has at least five equivalents in the Swedish: "snöglopp", "snöyra", "modd", "snösörja" and "kram" and many more in the Eskimo language. Categorisation and terminology of colours is an endless (therefore ideal) research field in relation to various theories of categorisation. Grammatical categories are also full of differences: e.g. the disparity between the category of grammatical gender in Polish and English – the latter one having this distinction only with relation to living beings. The language itself is also part of the world. When we speak of nouns, verbs, phonemes, sentences we are also categorising. While the issues whether the structure of the language influences the perception and cognition of the world, including categorisation, and whether cognition determines language facts remain open.

Categorisation is an indispensable attribute of cognition. By noticing and appreciating similarities among potentially diverse objects one is able to extend generalisations based on previous experiences. Vis-à-vis the complexity of the surrounding world, without the ability

* It is perhaps worth reminding the reader here that in the past subject heading systems have been known as subject classifications.

[3] Woźniak J. *Kategoryzacja: studium z teorii języków informacyjno-wyszukiwawczych*. – Warszawa: SBP, 2000.

** Simplifying the matter one can say that the classical attitude to categorisation is based on the four following assumptions: a) categories are defined by necessary and sufficient features; b) features are binary; c) categories have distinct limits; d) all elements of a given category are equivalent.

of recognizing similarities, which exist among otherwise dissimilar objects, we would by at a complete loss. Hence, categorisation is the means of simplifying reality, reducing the excessive dependence on memory, and a measure aimed at effective collecting and retrieving information, etc. Categories, just like classes in classifications, cannot exist in isolation, but are joined by relations, which encompass particular classes within superior, more general categories.

Categorisation in indexing languages

The term "category" was introduced into the theory of indexing languages by Shiyali Ramamrita Ranganathan, who distinguished five categories: individuum, matter, energy, space and time. An analysis of the uses of this term demonstrates that in the most general sense it is understood as a class of objects belonging to a certain universum. This class is called into being by a division based on features, which, from the point of view of criteria of relevancy, are considered elementary. Categories, being indices of very general connotations, to which one can bring other meanings, e.g. by way of defining, indicate the main genres or types of objects of the categorised universum. Concepts of a high degree of generality with a wide area of application elaborated by the mind in referring directly or indirectly to empirical knowledge and utilized by the mind in interpreting such knowledge. Categories are used for grouping expressions usually by relating fundamental aspects of reality. Some of the categories defined in this way demonstrate ontological character. Categorisation the words of the indexing languages is another issue.

To exploit fully categorisation as a method of optimisation of language structure organisation requires among other good orientation in the reality described by the language and of the relevance criteria, which are employed by the users. In everyday practice however, for numerous reasons subject heading systems are seldom designed on the basis of such knowledge. A drastic example in this instance can be the application of academic subject heading systems (which reflect the scientific view and way of approaching reality) to non-scientific information collections, gathered e.g. by public libraries.

Information organisation models encountered on the Internet are other additional factors enhancing the "popularity" of categorisation. Catalogues of subject categories aimed at helping the user in browsing the networks, are an instance of the category approach oriented at the user. Such catalogues are hierarchically structured sets of classes generated empirically on the basis of initial division of the holdings of the Internet into tens of general classes relative to the most commonly searched types of information. Hence, alongside the class called "Science" one can find "News", "Computer and Internet" or "Shopping". Subject categories of various degree of development belong nowadays to the most popular methods of information organising in numerous types of Internet services. They are employed both in great portals of universal type, specialised gateways, and in corporate websites.

Relational lists

The most complex organisational models are found in the class called relational lists, which comprises among other: thesauri[4] and information ontologies. Numerous thesauri used in the network environment are controlled vocabularies, usually elaborated for the needs of detailed disciplines and practical operations. Many of the thesauri currently under

[4] Woźniak-Kasperek J. *Podstawy budowy tezaurusa: poradnik*. – Warszawa: SBP, 2006.

construction do not meet required norms, violate standards, sometimes they are not thesauri at all, but simple alphabetical lists of keywords. In most cases such situations result form the fact that they are designed by persons with not enough necessary knowledge and competence.

The last few years have witnessed intensive development of research and implementation of the so called ontologies. An ontology is a type of description of expressions and relations, which are to be recognisable for computer programs. In other words, an ontology can be defined as a set of formalised expressions, defining relations between expressions and setting the rules of reasoning. The number of semantically rich methods of approaching ontology modelling is constantly growing. Computers are (or more precisely, will be) able to understand the semantic content of internet documents, drawing upon ontologies, to which the expression used in these documents refer. It seems that the importance of ontologies in the systems of knowledge management and in the future Semantic Network will be increasing. Already today ontologies are used in e.g. describing products and services in electronic commerce, or in museology (in digital descriptions of the organisation of museum collections).

Conclusion

Internet has provided the individual user and libraries with access to constantly increasing resources of digital information. Librarians are challenged with the task of working out new methods of helping the users in finding the necessary information. The ideas of knowledge society, continual education, the growing tendency to shop on the web, the concept of citizen-authorities contact through electronic media etc. render access to information, i.e. the ability to research, select and evaluate data, increasingly important. Indexing and retrieving are mutually conditioning procedures. If an object is methodologically described then the search engine furnished with a similar methodological mechanism is capable of finding it. If users are to find data effectively in diffused collections, the metadata which describe the objects searched for and their contents have to be cohesive with some sort of a coding model or a controlled vocabulary system. Indexing languages which use controlled vocabulary (or as they are frequently called among Internet context: controlled vocabulary systems) are (or should be) the tool which supplements keywords. Several years ago I would have phrased it: keywords supplement researching with controlled vocabulary. Today, it seems to me the most appropriate to say, that keywords and controlled vocabulary supplement one another, in other words are complementary. Languages with controlled vocabulary have just as many advantages as flaws. Keywords also; although in this case we are discussing different sets of advantages and flaws. One could somewhat lightly say, that controlled vocabulary languages and keywords resemble democracy. We can compare them and argue (sometimes with too much conviction) the superiority of one vis-à-vis the other; we criticise them vehemently, but in the end we use them, not because they are good, but because at the moment there is nothing better at hand. Perhaps the future belongs to intelligent machines which will interact similarly to humans. But awaiting the future with hope (and some anxiety, at least on my part) we must learn how to use better what we have now at our disposition.

ОРГАНІЗАЦІЯ ІНФОРМАЦІЇ У ПРЕДМЕТНИХ ШЛЮЗАХ

Ядвіга ВОЗНЯК-КАСПЕРЕК

Варшавський університет, Інститут наукової інформації та книгознавства,
вул. Нови Свят, 69, м. Варшава, 00-046, Польща,
ел. пошта: jbwozniak@uw.edu.pl

Стаття присвячена одній із ключових проблем сучасного суспільства – проблемі оптимізації пошуку інформації. Автор розглядає питання участі експертів у підборі й індексації інформаційних об'єктів і важливість їхнього внеску. Двома додатковими чинниками у розв'язанні проблеми представлення і пошуку інформації є контрольовані словники і неконтрольовані набори ключових слів.

Ключові слова: пошук інформації, контрольовані словники, системи організації знання, предметні шлюзи.

ОРГАНИЗАЦИЯ ИНФОРМАЦИИ В ПРЕДМЕТНЫХ ШЛЮЗАХ

Ядвига ВОЗНЯК-КАСПЕРЕК

Варшавский университет, Институт научной информации и книговедения,
ул. Новы Свят, 69, г. Варшава, 00-046, Польша,
эл. почта: jbwozniak@uw.edu.pl

Статья посвящена одной из ключевых проблем современного общества – проблеме оптимизации поиска информации. Автор рассматривает вопрос участия экспертов в подборе и индексации информационных объектов и важность их вклада. Двумя дополнительными факторами в разрешении проблемы представления и поиска информации есть контролируемые словари и неконтролируемые наборы ключевых слов.

Ключевые слова: поиск информации, контролируемые словари, системы организации знания, предметне шлюзы.

Стаття надійшла до редколегії 29. 08. 2009
Прийнята до друку 16. 04. 2010