Utilizing Faceted Structures for Information Systems Design

Uta Priss, Elin Jacob
School of Library and Information Science,
Indiana University Bloomington

Abstract

Even for the experienced information professional, designing an efficient multi-purpose information access structure can be a very difficult task. This paper argues for the use of a faceted thesaurus as the basis for organizing a small-scale institutional website. We contend that a faceted approach to knowledge organization can make the process of organization less random and more manageable. We begin by reporting on an informal survey of three institutional websites. This study underscores the problems of organization that can impact access to information. We then formalize the terminology of faceted thesauri and demonstrate its application with several examples.

Analyses of Website Structures

Evaluation of the organizational structure of a website must necessarily involve an assessment of its hyperlink structure and of the support the linking structure provides for efficient access to information. Working from the basic assumption that a well-structured website will improve access, websites of three leading Library and Information Science [LIS] were examined to determine the manner in which the hyperlink structure of each supported or hindered access to information. Because the purpose of this exercise was to determine the potential for a more structured approach to website design, it was decided to conduct an informal analysis of each site rather than to undertake a time-consuming formal study.

Review of the site map provided by each website revealed different approaches to overall organizational structure. While one sitemap indicated that pages were grouped according to conceptional content, another seemed to be based on a mixture of conceptual and usage-based organization. The third website indicated that the web pages were only loosely grouped.

The organization of each site was informally evaluated by determining the efficiency with which the following questions could be answered:

- 1) How is the school ranked in national evaluations?
- 2) Are there members of the faculty who are experts in the field of knowledge organization?

- 3) Is it possible to obtain a specialized degree, such as music librarianship, from this school?
- 4) When does the Fall/Spring semester start?
- 5) What journals are published by the school or edited by faculty members?

It was decided in advance that each of the five questions must be answered in less than three minutes. Furthermore, assessment of website efficiency in addressing a particular question would be based on the number of mouse clicks required to access requisite information.

The first question reflects how temporary information is handled. Because all three schools were recently ranked in a national survey, it was expected that this information might appear under ”news.” This is the case for one website, which may actually discard the information as soon as the news becomes ”old”. Another website has a separate link labeled ”ranking”; and the third website provides access by selecting ”general information” and then ”about the school”. Thus, while one website appears to consider this information temporary; the other two seem to consider it permanently relevant.

The question regarding specific research interests examines problems with the polyhierarchical nature of information access structures. The topic ”faculty members according to research area and/or subject expertise” can be grouped under the two broader headings, ”research” or ”faculty”. While the answer to the question can be retrieved from all three websites, only one website lists the topic under both broader topics, although one of the links is not correctly implemented and does not point to the appropriate page. Another website maintains nearly identical versions of the ”faculty by subject expertise” page in two separate locations. While the design of these two pages differs significantly – indicating, perhaps, that they are actually maintained by different people – the information provided on each page is very similar. On the third website, the links under ”research” do not lead to the faculty listing. If the user begins with ”research”, she must ultimately return to the first page and select ”people” and then ”faculty by subject”.

The question about specialized degrees was formulated to address information that would not appear at the upper levels of the website. For only one website is a definite ”no” provided. For the other two, the answer is not clear. In one case, the user must intuit that this information would be available under the heading ”catalog”. With each of the websites, however, information about degrees can be obtained by scrolling through lengthy bodies of text detailing academic regulations.

Because the university calendar is maintained by the university administration, the question about semester dates addresses information for which a school may not feel responsible. One website provides access to three separate calendars in three different locations: a calendar of events without semester dates maintained by the school; admission pages with deadlines for registration and dates for orientation events; and the academic calendar maintained by the registrar’s office which is reached only through

several links beginning with ”course catalog”. The second website has a complete calendar that contains academic information as well as events and holidays and includes links to other academic calendars. The third website also maintains a calendar that is directly accessible from the first page.

The final question regarding in-house publications is concerned with information that may not be relevant for every school. Users will not know whether a non-existing link to in-house publications implies that the publications do not exist or whether the link is simply missing. The question was easily answered by one website which links to its publications office from the first page; but it could not be answered for the other two websites within the three-minute limit. This information might be accessible under ”research centers”, but it would take time to check each individual center. And, while one website provides information about a technical report series under the heading ”research”, it is not clear if this is the only publication edited by that school.

All in all, the three websites reviewed in this informal and very limited survey appear to be roughly comparable with respect to ease of access through navigation of the website’s hyperlink structure. For example, the website with the most structured site map has problems with implementation on lower levels. In contrast, the website that claims only a loose grouping of pages is fairly efficient in addressing most questions. It is interesting that two of the websites make all first and second level links available on the first page, either via a frame or via a Javascript implementation that shows links as ”mouse over” events. For both of these websites, it generally takes only one or two mouse clicks to access the desired information. The third website uses a more traditional hierarchical structure that moves from directory to directory and consequently requires more mouse clicks and time to navigate.

A Faceted Thesaurus as a Formal Model

This informal review of website organization indicates how difficult it can be, even for the experienced information professional, to design an effective and efficient multipurpose access structure. We argue that application of a faceted approach to knowledge organization can ensure that the process of organization is less random and more manageable. To this end a faceted thesaurus is recommended as the basis for a systematic approach to structuring a website. An artificial intelligence knowledge base or a lexical database, such as Princeton’s WordNet would be equivalent to a non-faceted thesaurus since they contain a generic relation [BTG/NTG], a part-whole relation [BTP/NTP], and other relations, such as a related term [RT] relation. It should be noted that ”generic relation” is adapted from ISO’s ”generic relationship” (ISO 2788-1986). Mathematically, a relation is comprised of relation elements: for example, the set of all BT/NT relationships in a thesaurus forms the generic relation. Single pairs of BT/NT relationships are elements of the relation. In a larger non-faceted thesaurus, it can be difficult to maintain coherence and consistency across these relationships. The adaptability and flexibility of a faceted thesaurus make it decidedly preferable as an aid to the structuring process of website design.

The formal definition of a faceted thesaurus has been revisited and re-defined to allow implementation in an object-oriented environment. The resulting notion of ”thesaurus” is, on the surface, very similar to traditional conceptions of thesauri (compare Soergel (1985)). On a deeper level, however, the reformulated definition of thesaurus provides a simple and mathematically consistent model that allows a precise description of the underlying conceptual theory and can be directly translated into an object-oriented representation.

A thesaurus consists of terms (or descriptors) that are arranged in a generic relation. Further relations, such as a part-whole relation and a ”related term” relation are defined among the set of terms. The generic relation need not form a tree hierarchy because it is assumed that the relation is (or can form) a polyhierarchy or, to use the mathematical term, an ordered set. This convention is not a restriction because the notion of an ordered set is broader than a tree hierarchy, which is simply a special kind of ordered set.

Terms are defined conceptually, not linguistically. This means that a ”single term” can be a word or a phrase or a word followed by a scope note (a disambiguating remark in parentheses), etc. Terms are to be distinguished from concepts. Each term describes a concept. Several terms are considered synonymous if they describe the same concept. However, if the faceted thesaurus is to be used for structuring a website and not as an end-user product, it is not necessary to account for synonyms. It can therefore be assumed that single terms are not synonymous and that each term in the system is a preferred term. This is not a restriction on modeling since synonyms can be added if required.

Although every single term describes a concept, not every concept can be described by a single term. In addition to single terms, composed terms – terms generated by the combination of single terms from different facets – are also used to describe concepts. Term composition is to be distinguished from term aggregation. Composed terms form a concept, such as ”white rose”, whereas terms which are simply aggregated, such as ”male” and ”female”, can co-occur in a document or on a web page but do not form a concept ”male female”. Any term can be aggregated with any other term because potentially any two topics can co-occur in a document. But term composition is constrained by rules: for example, mutually exclusive terms, such as ”male” and ”female” cannot normally be composed. However, facets can be constructed either by term aggregation (”male” and ”female” can be aggregated in a facet ”gender”) or by term composition (”gender” and ”species” can be composed in a facet ”animal” resulting in concepts such as ”female feline”).

The distinction between ”aggregation” and ”composition” is adapted from De Morgan (1860) and is essential for this formalism. Terms and concepts form separate ordered sets (hierarchies). As will be shown below, in case of term aggregation the resulting term hierarchy is essentially identical to the concept hierarchy whereas in case of term composition the resulting hierarchies are different. While the term hierarchy (or the generic relation on the set of terms) constitutes the enumerated backbone of the

```
faculty member
      permanent (faculty)
      adjunct (faculty)
      visiting (faculty)
      emeritus (faculty)
```

Figure 1: A baseline facet

thesaurus, the concept hierarchy is dynamically created in runtime. Therefore only the term hierarchy needs to be stored in a database. The set of facets itself also forms an ordered set. This ensures that no facets are redundant (i.e., occur more than once); that all facets can be designed in a modular fashion; that there are no cycles in the structure; and that consistency and coherence of relations can be maintained. Each hierarchy can be displayed: textual displays show the enumerated term hierarchy; facet hierarchy diagrams show the hierarchical relation among facets; concept diagrams show the dynamically created concept hierarchy; and relationship diagrams visualize relationships of terms in a manner similar to entity-relationship diagrams of relational databases.

The minimal elements of a thesaurus are baseline facets, which are defined as follows.

Definition 1:
A *baseline facet* consists of a set of terms, a set of concepts, and a generic relation on the set of terms with the following conditions:
a) the generic relation is an order relation on the set of terms (mathematically, a relation that is defined as reflexive, antisymmetric and transitive);
b) the ordered set of terms has one top element (called the *head term*) which is broader than any other term in the baseline facet;
c) there is a one-to-one mapping from the set of terms to the set of concepts (i.e., every term describes a concept, no two terms describe the same concept and vice versa).

Fig. 1 shows an example of a baseline facet. The generic relation is indicated by indenting the narrower terms. For baseline facets, the concept hierarchy is identical to the term hierarchy and could be transformed into a concept lattice in the sense of formal concept analysis (Ganter & Wille, 1999). Baseline facets can be more complex than the example in Fig. 1. More complex facets, such as the facet in Fig. 2, can be baseline facets, but these more complex facets are frequently constructed by combination of less complex facets.

The next definition explains several types of facet construction. Examples for each type are given further below.

Definition 2:
a) *Facet construction by term aggregation*: the set of terms of the resulting facet is equal to the union of the terms of two or more component facets which is joined

with a new head term. The resulting set of concepts is the union of the concepts of the component facets joined with a new top concept. The resulting generic relation is the union of the generic relations of the component facets joined with the relation between all terms and the new head term. The concept hierarchy is defined such that a concept is a subconcept of another concept if their terms are in generic relation to each other.
b) *Facet construction by term composition*: the set of terms of the resulting facet is equal to the union of the terms of two or more component facets which is joined with a new head term. The resulting generic relation (on the set of terms) is the union of the generic relations of the component facets joined with the relation between all terms and the new head term. The resulting set of concepts is the direct (Cartesian) product of the sets of concepts of the component facets multiplied by the new top concept. The concept hierarchy is defined such that a concept is a subconcept of another concept if the subconcept relation holds for their component concepts.
c) *Facet construction by term composition with constraints*: the conditions of term composition are fulfilled but the set of concepts of the resulting facet is not the full direct product of all sets of concepts of the component facets. The constraints prescribe which sets of concepts are combined in the direct product (or in several products), the results of which are joined.

Because a unique head term is required for every new facet it follows that every facet can be identified by its head term. For example, the baseline facet in Fig. 1 can be identified as the ”faculty member” facet. The following concepts are defined for all types of facet construction.

Definition 3:
a) A facet is called a *subfacet* of another facet if it is used as a component in constructing the facet.
b) A term is called *element* of a facet if it is an element of the set of terms in that facet.
c) A *facet hierarchy* is a set of facets with a subfacet/facet relation which forms an ordered set and for which, for each term, there is exactly one minimal facet so that the term is an element of that facet.

Fig. 2 contains an example of facet construction by term aggregation where the terms of the baseline facets ”faculty”, ”student” and ”staff” are joined to form a new facet for which the term ”people” is chosen as head term. Occasionally, head terms may be changed when creating a new facet. For example, initially the ”faculty” facet might have been called ”people”. When it is aggregated with the ”student” and ”staff” facets, the names may be changed so that ”people” becomes the new head term of the combined facets. Name changes do not pose a problem as long as there is a thorough check to verify that names are not duplicated and that all occurrences of a name are changed simultaneously. Because term aggregation is the default method of facet creation it carries no special symbol (such as NT). Instead, the generic relation is indicated by indentation of narrower terms. The notational symbol behind ”student” indicates that ”student” is the head term of a baseline facet that is defined somewhere else (Fig. 4). For facets created by term aggregation, the concept hierarchy corresponds to the term

```
people
       faculty member
                regular faculty
                adjunct faculty
                visiting faculty
                emeritus faculty
       student @
       staff
                administrative staff
                technical staff
                library staff
```

Figure 2: A facet created by term aggregation

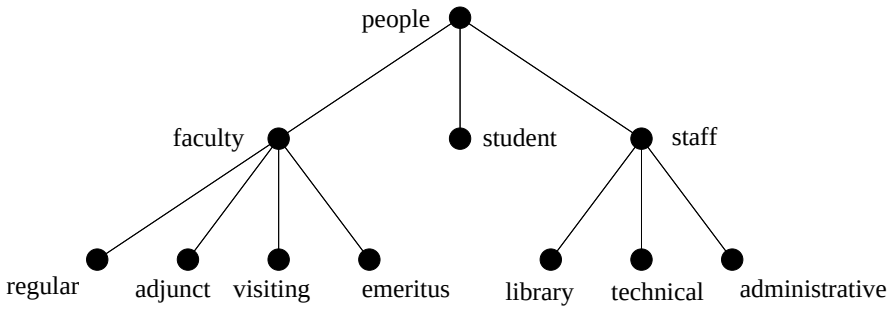


Figure 3: The concept diagram of the facet in Fig. 2

hierarchy. Fig. 3 shows the concept diagram of the corresponding concept hierarchy for the term hierarchy in Fig. 2.

Facet construction by term composition is the most significant feature of a faceted thesaurus. If two facets are combined term composition, the set of concepts results in all possible combinations of concepts among the facets. These combinations need not be enumerated. Fig. 4 gives an example. The head terms of the single facets are included in angle brackets to denote facet construction by term composition.

All concepts under ”by enrollment status” can be combined with all concepts under ”by residence”. And they can all be combined with ”current student”. On the other hand, ”by enrollment status” and ”by residence” are not useful for ”alumnus”. Therefore, in this case, term composition with constraints might be more appropriate. The constraints should be listed following the head terms of the subfacets: for example, ”by enrollment status: applies only to student by type, current”.

Fig. 5 shows the corresponding concept hierarchy (by composition with constraints). The upper half of the figure shows a regular display, the lower half a so-called ”nested line diagram” as used in formal concept analysis (Ganter & Wille, 1999). Although

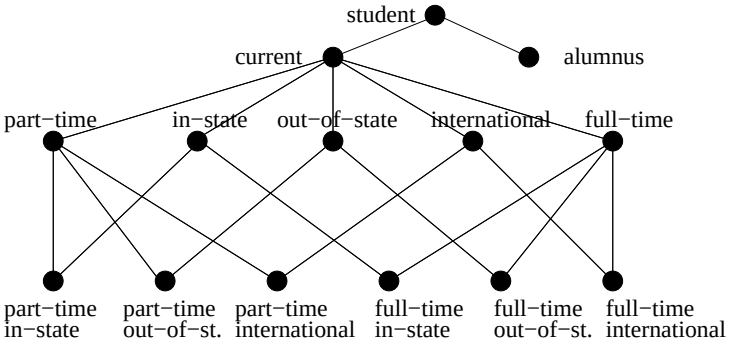
```
student
        <by type>
                current (student)
                prospective (student)
                alumnus
        <by enrollment status>
                part-time
                full-time
        <by residence>
                in-state
                out-of-state
                international
```

Figure 4: A facet created by term composition

the nested line diagram reduces the complexity, it is obvious that concept hierarchies can easily become very complex. It is therefore an advantage of the faceted thesaurus that concept hierarchies are not stored in the system but are generated as needed. The process of term composition does not imply any ordering of the composed terms: ”in-state/part-time” is identical to ”part-time/in-state”. This is in contrast to classification schemes where a prescribed citation order determines the order in which facets are combined. The nested line diagram requires a citation order because one facet (here ”by enrollment status”) must serve as the outer structure, while the other facet (”by residence”) serves as the inner structure. In a computer-generated representation the citation order would be flexible, just as formal concept analysis software allows users to alter the citation order.

Definition 3c ensures the coherence and consistency of a faceted thesaurus. Each facet has a precise place in the facet hierarchy. Because the facet hierarchy is an ordered set, each facet can be subfacet of multiple (super-)facets. But Definition 3c prohibits facet A as a subfacet of facet B if facet B is a subfacet of facet A. A graphical representation of the facet hierarchy, which is comparable to the file manager software on a PC, is used to control and maintain the hierarchy and to locate facets. Fig. 6 shows an example of a facet hierarchy. The minimal elements in the diagram are baseline facets, whose lower-level terms are not included in the diagram. Definition 3c also ensures that every term has a precise location in the facet hierarchy and that most terms occur in exactly one baseline facet. Those terms which do not occur in a baseline facet are head terms of higher-level facets.

The part-whole relation and related term relation can be modeled as types of facet construction by term composition. For example, an organization has ”chair”, ”chapter”, ”committee” and so on as its parts. Although these are mutually exclusive because an entity cannot be a chair and a chapter at the same time, they can be combined with respect to the concept ”organization”. An organization may have at least one chair,



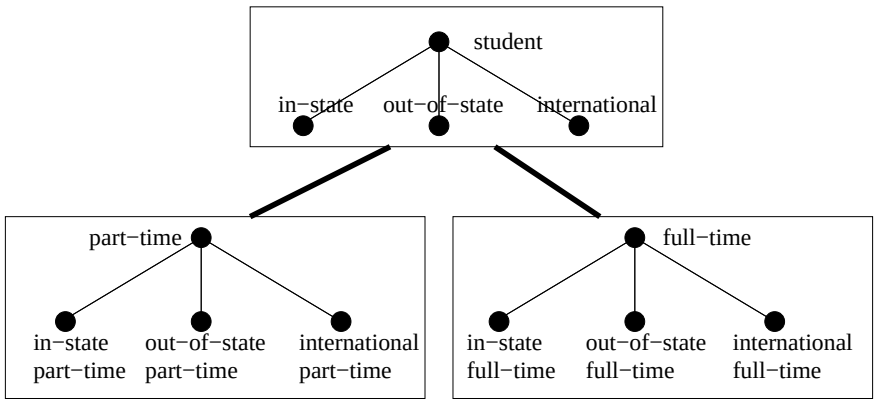


Figure 5: Two concept diagrams of the facet in Fig. 4

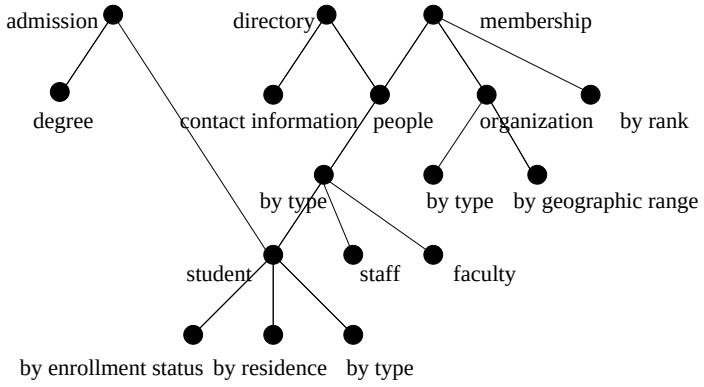


Figure 6: A facet hierarchy

```
organization
        <by type>
               academic (organization)
               professional (organization)
        <by geographic range>
               local
               state
               region
               national
               international
NTP
        <chair>
        <chapter>
        <SIG>
        <committee>
```

Figure 7: A facet with NTPs

chapter, committee and so on. Therefore, parts are similar to facets that are combined by term composition. Each part is enclosed in angle brackets and the list of parts is preceded by the notation ”NTP”, as shown in Fig. 7.

In contrast to the traditional modeling of related terms as a symmetric relation, the related-term relationships of a faceted thesaurus are similar to relations in a relational database, which are presented as tables. If two terms (or facets represented by their head terms) are related, then a new facet is defined based on term composition of the two terms. Although this process is a type of facet construction by term composition, there is a conceptual difference between forming facets through term composition and forming facets that represent a relationship. Therefore, the head terms of the relational subfacets are denoted by double angle brackets. As indicated in Fig. 8, ”membership” is a relation between ”people” and ”organizations”. Instead of implementing ”people RT organization” and vice versa, a new facet ”membership” is created that has ”people” and ”organization” as relational subfacets. The relationship can be formed of as many subfacets as desired. Other subfacets that do not constitute the relationship but are subfacets of the resulting facet, such as ”rank” in this example, can be added.

Although they are structurally similar, related terms and part-whole relations are distinguished from term composition because they are conceptually different. Inverse relations, such as BT, BTP and BTR can be computed at runtime and added to the display. Term composition, part-whole relations and related term relations are inherited by subfacets. Therefore parts and relationships should be implemented between the broadest possible facets. For example, although the parts that are listed under ”organization” are inherited by both ”professional organization” and ”academic organization” and can be represented as such in a display, the parts are enumerated only once in the formal structure.

```
membership
        <<people @>>
        <<organization @>>
        <by rank>
                associate
                full
                probation
```

Figure 8: A relational facet

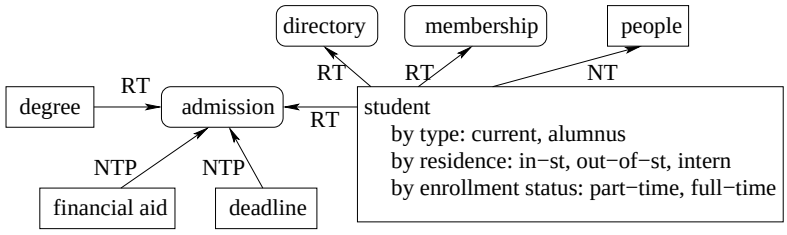


Figure 9: A relationship diagram

IMPLEMENTING AND USING A FACETED THESAURUS

Because it would be virtually impossible for human users to keep track of such a complex structure, access to and maintenance of the formalized structure should be provided by a software tool that holds to Definitions 1, 2 and 3. The terms, relationships and facets of a faceted thesaurus are stored in a database. Access to the faceted thesaurus is provided by four types of displays: text displays (Figs. 1, 2, 4, 7 and 8); facet hierarchy diagrams (Fig. 6) that show the facet/subfacet relation; relationship diagrams (Fig. 9) that show the immediate relationships of selected terms; and concept diagrams (Figs. 3 and 5). The diagram in Fig. 9 visualizes the relationship between ”student” and ”admission”. For both terms, the immediately related sub- and superfacets are also displayed. Narrower terms have been omitted except for ”student”. Users can customize the diagrams by choosing how many levels they want to see, whether narrower terms are to be included, whether arcs are to be labeled, and so on.

The different graphical representations allow a modular design because relationship diagrams and concept diagrams only show the immediate neighborhood of a term. The displays facilitate top-down strategies (facet hierarchy and relationship diagrams for high-level facets) and bottom-up strategies (relationship diagrams for low-level facets). Although the formalism implies a bottom-up conceptual structure, this is not mandatory for the actual design process. High-level relationship diagrams can be constructed first and then filled with lower-level facets and terms.

A faceted thesaurus provides a controlled vocabulary of descriptors and relationships that can be used to represent the conceptual content of a document collection.

In a website application, appropriate descriptors are included in the metadata of each webpage. The hyperlink structure of the site can then be created interactively from the relationship diagrams that depend upon the metadata descriptors. The coherence and consistency of the website is ensured because every facet has a precise location in the thesaurus. As a result, each webpage has a precise location within the hyperlink structure. The flexibility of the representational structure facilitates implementations that can respond to the needs of specific user groups by combining website content facets with a user need facet. High-level relational facets can then provide ”walking tours” through the website for new users and customized access points for current users, prospective users and administrators.

Concerning the five questions that were examined in the first section, it can be summarized that a polyhierarchical organization can easily be maintained (question 2). Before a new page is added to the site, the facet hierarchy is checked to identify related pages and to prevent duplication of pages. New hyperlinks are adjusted according to the facet hierarchy. Once the faceted thesaurus is created, it will be fairly stable and because of the dynamically created concepts it will be able to accommodate new topics. The ”news” page (question 1) can be compiled as a weekly updated relational facet that draws from topics in the rest of the thesaurus. All relevant topics can be maintained in the thesaurus, not only the top-level concepts (question 3). Similarly to a ”user needs facet”, an ”authorship” facet can coordinate information from webpages that are maintained by different institutions (question 4). The faceted thesaurus can be personalized for an application (question 5) by creating appropriate top-level facets.

REFERENCES

De Morgan, A. (1860). *On the syllogism, no. IV, and the logic of relations.* Trans. Cambridge Phil. Soc., 10: 331-358.

Ganter, Bernhard; Wille, Rudolf (1999). *Formal Concept Analysis: Mathematical Foundations*. Springer Verlag, Berlin-Heidelberg.

ISO 2788-1986. International Standard 2788: *Documentation -Guidelines for the establishment and development of monolingual thesauri*. International Organization for Standardization.

Soergel, Dagobert (1985). *Organizing information*. San Diego, CA: Academic Press.