A Taxonomy of Lexical Metadata Categories

Bodil Nistrup Madsen, Hanne Erdman Thomsen
Department of International Language Studies and Knowledge Technology,
Copenhagen Business School
Dalgas Have 15, DK-2000 Frederiksberg, Denmark
E-mail: bnm.isv@cbs.dk, het.isv@cbs.dk

Abstract

Metadata registries comprising sets of categories to be used in data collections exist in many fields. The purpose of a metadata registry is to facilitate data exchange and interoperability within a domain, and registries often contain definitions and examples. In this paper we will argue that in order to ensure completeness, consistency, user-friendliness and extensibility, metadata registries should be structured as taxonomies. Furthermore we will illustrate the usefulness of using terminological ontologies as the basis for developing metadata taxonomies. In this connection we will discuss the principles of developing ontologies and the differences between taxonomies and ontologies. The paper includes examples of initiatives for developing metadata standards within the field of language resources, more specifically lexical data categories, elaborated at international and national level. However, the principles that we introduce for the development of data category registries are relevant not only for metadata registries for lexical resources, but for all kinds of metadata registries.

1. Introduction

In order to facilitate data exchange and interoperability, it is important to be able to describe elements of data collections systematically and unambiguously. This is the reason why metadata registries comprising sets of metadata categories, giving accepted definitions and examples, exist in many fields.

A given set of metadata categories may be used not only for the description of data elements with a view to obtaining a common understanding of data elements and to facilitating data exchange, but also as a guideline for which data categories to choose when creating a new data collection. When defining a set of metadata categories it is very useful to base it on a kind of systematization, e.g. a taxonomy, specifying main categories, categories and subcategories. Otherwise one may end up with an incomplete and inconsistent set of categories that is very difficult to use and to extend.

Metadata registries are used whenever data must be used consistently within an organization or group of organizations. Therefore a huge number of metadata registries have already been developed. Very often it is, however, only possible to obtain an alphabetic listing of the data categories, cf. for example the Metadata Online Registry for national data standards for health, housing and community services statistics and information of the Australian Government, METeOR.

In this paper we will give an introduction to some existing metadata initiatives and propose principles for a systematization of metadata categories that will meet the above mentioned requirements for completeness, consistency, user-friendliness and extensibility. We will argue that metadata taxonomies should always build on the principles of creating terminological ontologies (concept systems), cf. for example ISO 704:2000 and Madsen, Thomsen and Vikner (2004).

Here we will present two metadata initiatives within the field of language resources: an initiative of the International Standardization Organization, ISO, and an initiative of the Danish Standards Association, DS.

2. Data Categories for Language Resources

ISO Technical Committee 37, Terminology and Other Language Applications, ISO TC 37, published a standard in 1999 specifying data categories used in terminological resources, ISO 12620:1999, *Computer assisted terminology management — Data Categories*. This standard was prepared by TC 37 Sub Committee 3, at that time having the title Computer applications in terminology. In 2003, TC 37/SC 3 initiated a revision of the existing document with the intention of creating a family of data category standards designed to meet the needs of terminologists and other language experts developing a variety of electronic linguistic resources, cf. Wright (2004). The intention was to include data categories for a variety of applications, including for example terminological and lexicographical data collections as well as machine translation lexica. Standards for these three kinds of data collections are developed in three sub committees of TC 37: SC 3 (Systems to manage terminology, knowledge and content), SC 2 (Terminographical and lexicographical working methods) and SC 4 (Language resource management). At the same time it was suggested to set up a Data Category Registry (DCR) for all the above mentioned kinds of lexical data, cf. also Ide & Romary (2004). The DCR is intended to be compliant with ISO 11179-3, *Information technology — Metadata registries (MDR) — Part 3: Registry metamodel and basic attributes.*

In the following two sections we will first describe the structure of the original standard for data categories in terminological data collections, ISO 12620:1999, and then discuss the proposed structure for the DCR.

2.1 The structure of ISO 12620:1999

The data categories of ISO 12620:1999 were classified in

three major groups: *data categories for terms and term-related information*, *descriptive data*, and *administrative data*. The groups were further subdivided into ten sub-groups:

Term and term-related data categories:
- A.1 term
- A.2 term-related information
- A.3 equivalence

Descriptive data categories:
- A.4 subject field
- A.5 concept-related description
- A.6 concept relation
- A.7 conceptual structures
- A.8 note

Administrative data categories:
- A.9 documentary language
- A.10 administrative information

This structure is not homogenous, i.e. it reflects various subdividing criteria (dimensions), and it does not give a very clear overview of the data categories. One dimension is for example term-related information vs. concept-related description. Here it is not clear why e.g. *subject field* and *concept relation* do not fall within the group: concept-related description.

An example of term-related information is ***A.2.1.18.1 collocation***, while an example of concept-related information is ***A.5.3 context*** (a text or part of a text in which a term occurs). Types of *contexts* include:

a) *defining context*: a context that contains substantial information about a concept, but that does not possess the formal rigor of a definition

b) *explanatory context*: a context that provides a summary explanation of a concept

c) *associative context*: a context that contains the minimum amount of conceptual information needed to associate a concept to a particular concept field

d) *linguistic context*: context that illustrates the function of a term in discourse, but that provides no conceptual information.

It is not clear why *linguistic context* is categorized as concept-related information, c.f. the explanation in *d)*.

It seems as if the structure of ISO 12620:1999 is to some extent based on the structure typically found in a terminological entry. Since the above mentioned DCR of TC 37 will also include data categories of dictionaries, this structure is not very appropriate. Wright (2004) says that this classification was difficult to arrive at and does not satisfy anyone. Consequently it was decided to give up a classification of the categories. In our opinion it will, however, be difficult to ensure completeness, consistency, user-friendliness and extensibility of the above mentioned DCR, if there is no structure of the data categories. We will come back to this in our proposal for structuring the DCR in section 4.

2.2 The structure of the DCR

As already mentioned, the DCR will contain data categories that are relevant in various areas, such as terminology, lexicography and machine translation. These areas are referred to as thematic domains (TD). As illustrated in Figure 1 (from Wright 2004), the various data category selections (DCSs), i.e. the subsets of the DCR corresponding to thematic domains, will overlap. For example, the data categories ***part of speech*** and ***grammatical gender*** will be relevant in all three of the thematic domains mentioned here.

This suggests that it will not be feasible to use the TDs as a basis for the structure of the DCR, and certainly it will not be possible to take over the structure of ISO 12620:1999, since it is specific to the TD terminology. Instead, a new structure should be introduced.

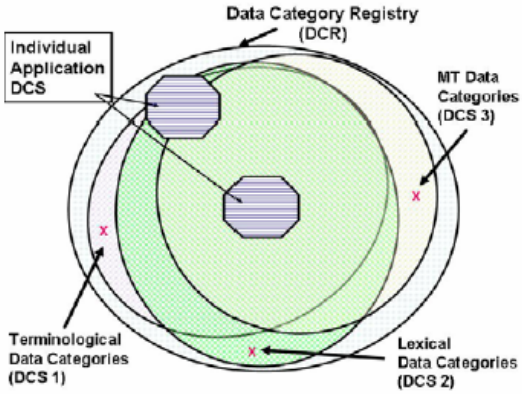

Figure 1: DCSs as subsets of the DCR

3. Principles for building Meta Data Taxonomies

In 2005, ISO TC 37/SC 3 decided to work on principles for building taxonomies for metadata, and it was stated that TC 37 should be a pioneer within this field, since this committee describes principles of concept modelling and classification in its standards and guidelines. On this background it was also argued that the DCR should be constructed as a taxonomy. Furthermore, it was argued that both data models and meta data taxonomies should be based on ontologies (concept systems), c.f. ISO TC 37/SC 3 N542 (2005).

In the next subsection we will introduce the concepts of ontology and taxonomy, and illustrate how ontologies can be used as a background for setting up a taxonomy.

3.1 Ontologies and Taxonomies

In recent years many authors have discussed the nature of ontologies and proposed various definitions and subtypes of ontologies for various purposes, among them Gruber (1993), Guarino (1998), Gómez-Pérez et al. (2004), Ruiz and Hilera (2006). However, the first two do not discuss the difference between ontologies and taxonomies, while the third set up a taxonomy of ontologies without discussing the difference.

According to the work done in the CEN Workshop, CEN

CWA 15045 (2004) *ontology* and *taxonomy* are types of knowledge structuring, as shown in Figure 2.

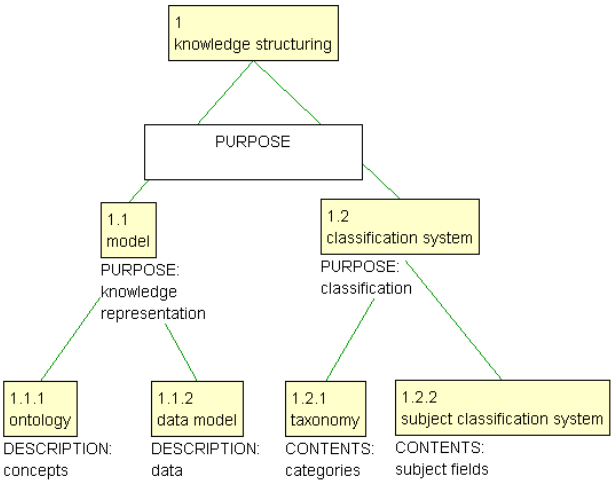

Figure 2: Ontology of knowledge representations

The ontology in Figure 2 comprises *concepts* (yellow boxes) and *subdivision criteria* (white boxes). The concepts are related by means of *type relations* (green lines) and further described by means of *feature specifications* each consisting of an *attribute-value pair* (e.g. *PURPOSE: representation of knowledge about phenomena*). The ontology in Figure 2 may be characterized as a *terminological ontology*, i.e. an ontology that is based on the terminological method, making use of characteristics and subdivision criteria.

According to the ontology in Figure 2, the purpose of a *model* is to give a simplified representation of knowledge about phenomena, whereas the purpose of a *classification system* is the subdivision of phenomena in classes.

In CEN CWA 15045 (2004) *ontology* and *taxonomy* are defined as follows:

> *ontology*: model of knowledge of the world comprising concepts and relations between concepts
> *taxonomy*: classification system for the classification of categories of a domain.

From the definitions and notes to definitions given in CEN CWA 15045, it can be deduced that classification systems typically comprise only type relations, whereas models include all kinds of relations, e.g. part-whole relations, temporal relations and causal relations.

In order to obtain a well structured taxonomy we will argue that it should be based on the elaboration of an ontology. In this way the concepts of the domain and their interrelations are clarified. In some cases it is even possible to 'generate' a taxonomy on the basis of an ontology, which means that the concepts of the ontology may more or less automatically be transformed into categories of a taxonomy. In other cases, the ontology may just render the knowledge, on which the construction of a taxonomy may be based.

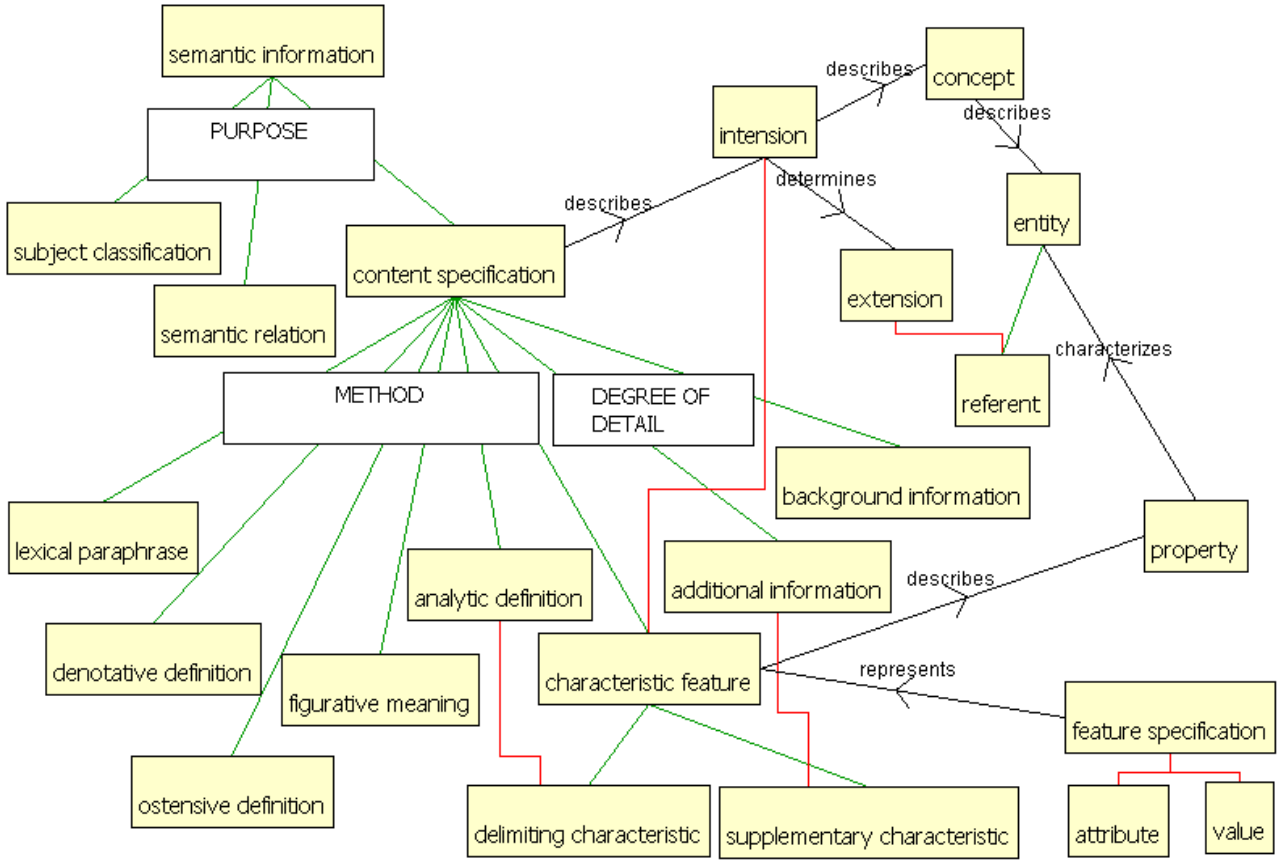

Figure 3: Ontology of semantic information

A taxonomy is often adjusted according to the intended purpose and user group. This means that some simplifications may be introduced into the taxonomy as compared to the ontology. In the next subsections we will give an example of an ontology and a taxonomy and we will describe some of the adjustments made to get from the ontology to the taxonomy. The example concerns lexical data collections.

3.2 An ontology of semantic information

Figure 3 presents an extract of an ontology for concepts pertaining to semantic information that may be registered in lexical data collections, such as e.g. termbases and electronic dictionaries.

The three main types of *semantic information* are *subject classification*, *content specification* and *semantic relation*.

This ontology comprises type relations, part whole relations (red broken lines) and associative relations (black lines with the designation of the relation type and an arrow indicating the direction of the relation).

The group of concepts on the right hand side, which are related by means of associative and part-whole relations, contribute to a better understanding of the concepts that are central for semantic information. For example, it is illustrated that a content specification describes the intension of a concept, and that the intension consists of characteristic features. At the same time *characteristic feature* is a kind of *content specification*. Also it may be seen that an analytic definition contains a delimiting characteristic. The full ontology of semantic information will also include for example the concepts *superordinate concept* and *subordinate concept*, and there will be a part-whole relation from *analytic definition* to *superordinate concept*.

3.3 A taxonomy: The Danish Standard of Lexical Data Categories

The Danish Standard DS 2394-1:1998 comprises a taxonomy for the classification of lexical data. This taxonomy is also referred to as the STANLEX taxonomy. The taxonomy was developed by a group of terminologists, lexicographers and computational linguists involved in machine translation and other kinds of natural language processing. It was developed in parallel with ISO 12620:1999 mainly because there was a need for a standard covering not only terminological data categories, but also data categories used for example in lexicographical data collections and in lexica of software for natural language processing. Consequently there was also a need for a systematic structure that was able to cover all these kinds of data collections. In STANLEX the main groups of information types are structured according to the main linguistic disciplines:

- etymological information
- grammatical information
- graphical information
- phonetic information
- semantic information
- usage

In addition to these categories there are categories for administrative information and structural information.

Examples of categories and sub categories are shown in Table 1. The entire taxonomy can be seen in the Appendix. All main groups, categories and subcategories are defined and exemplified in the standard.

Main group	Category	Subcategory
Semantic information	• Subject classification	• Classification system • Normative subject classification • Nonnormative subject classification
	• Semantic relations	• Concept system • Position of concept in concept system • Generic relation • Partitive relation • Successive relation • Causal relation • Associative relation • Antonymy • Metonymy • Equivalence within one language • Equivalence between two or more languages • Equivalence constraint
	• Content specification	• Lexical paraphrase • Analytic definition • Denotative definition • Ostensive definition • Additional information • Background information • Characteristic feature • Figurative meaning

Table 1: Categories and subcategories of Semantic Information

3.4 From Ontology to Taxonomy

The 'backbone' of the ontology in Figure 3 consists of the top concept *semantic information* and the subordinate concepts which are related to this concept by means of type relations: *lexical paraphrase*, *analytic definition* etc. These concepts will typically form the background for categories to be included in a taxonomy. As already mentioned, the concepts that are related by means of part-whole relations or associative relations typically give a better understanding of the central concepts, but it will often not be relevant to introduce corresponding categories in a taxonomy for lexical data collections.

The nodes in a taxonomy represent categories, not concepts, and a taxonomy category may sometimes represent more concepts. This may be more user friendly,

since the user of the taxonomy will then not have to think about subtle distinctions. For example one might decide to 'merge' the two concepts *additional information* and *background information* into one category in the STANLEX taxonomy, since it may be difficult for the user to choose between them. The concept *additional information* refers to information in the form of supplementary characteristics, while *background information* gives further information about historical, technical, legal or other aspects of the semantics of the lexical entry.

Sometimes the taxonomy will not comprise the 'lowest' levels of a hierarchy in the corresponding ontology. For example there may not be a need for distinguishing between *delimiting characteristics* and *supplementary characteristics* in the taxonomy. This is the case in the Danish Standard of lexical data categories.

In some cases it may be relevant to convert concepts of an ontology participating solely in associative or part-whole relations into categories in a taxonomy. For example it may be relevant to include the categories *feature specification*, *attribute* and *value* from Figure 3 as taxonomy categories.

4. Proposal for structuring the DCR

The structure of the STANLEX taxonomy gives a much clearer overview of the data categories than the original structure of ISO 12620:1999, and it is clearly better than a plain alphabetical list.

The use of a taxonomy for the structuring of a data category registry such as the DCR of ISO TC 37 makes it much easier to check whether the data category registry comprises all relevant data categories within a certain group. For example, in Table 2, that comprises categories and subcategories of Grammatical information, it can easily be identified that *countability* (for nouns in some languages) is missing. In this way the requirement for completeness of the data category registry is met.

In the case of proposals for new data categories it is also much easier to check whether the category is already in the DCR, maybe under another category name. An example of a data category name which is not transparent in ISO 12620:1999 is A.3.5 *transfer comment* that has the following explanation:

> *note included in a term entry providing more explicit information on the degree of equivalence, directionality or other special features affecting equivalence between a term in one language and another term in a second language*

If this subcategory belongs to a category of semantic relations that includes information on equivalence, it is much easier to identify than if it goes into an alphabetic list. Thus the taxonomic structure may prevent the maintenance authority of the data category registry from introducing doublet categories and in this way contribute to consistency of the DCR.

A metadata registry that is structured as a taxonomy may easily be extended in a consistent way. For example the structure and the definitions of the existing data categories make it easier to introduce new categories and subcategories that are clearly distinguished from the existing categories and subcategories. It may even be possible to introduce sub-subcategories. For example it may be possible to introduce *delimiting characteristics* and *supplementary characteristics* as sub-subcategories of the subcategory *characteristic feature* in the STANLEX taxonomy.

Main group	**Category**	**Subcategory**
Grammatical information	• Part of speech • Gender	
	• Information on inflection	• Stem • Paradigm information • Inflected form
	• Word formation	
	• Syntax	• Syntactic frame (valency) • Specification of syntactic frame • Specification of auxiliary verb • Syntactic function

Table 2: Categories and subcategories of Grammatical Information

The taxonomic structure and definitions based on this structure make it easier for a user of the DCR to find relevant data categories, since the user may not always be familiar with the names of the categories. It also facilitates the use of the metadata registry for choosing data categories when setting up a new data collection. In this way the structure contributes to user-friendliness.

On the background of the above mentioned advantages of using a taxonomy for the classification of metadata categories we suggest that the principles of the taxonomy of DS 2394-1:1998 are used for structuring the data categories in the DCR for lexical data in ISO TC 37. There will no doubt be a need for more categories and subcategories than those found in DS 2394-1:1998, but it will be easy to fit new categories into the structure, as long as they are mutually independent. There may also be a need for adjustments of the structure, since there do exist different ways of classifying lexical data. However, we think that DS 2394-1:1998 is a good starting point, and using the principles of this taxonomy will ensure completeness, consistency, user-friendliness and extensibility of the DCR.

The Registration Authority of the DCR should comprise members from all four sub committees of TC 37, and all Sub Committees, Working Groups and projects within TC 37 should report missing data categories, and should also comment upon addition or changes in existing data categories. It may also be feasible to add information about the applications of the individual data categories in the DCR.

5. Conclusion

In this paper we have compared examples of structures for metadata registries for lexical data collections. We have argued that by using a taxonomic structure, completeness, consistency, user-friendliness and extensibility are more easily obtained. Furthermore we have illustrated that the development of taxonomies should be based on terminological principles. Ideally the first step in the development of a taxonomy is to set up an ontology in order to clarify the relations and definitions of concepts that are central with a view to the taxonomy categories. In some cases the ontology may be mapped directly into the taxonomy, but in other cases it will be necessary and useful to introduce adjustments into the taxonomy compared to the ontology. The principles that we introduce here for the development of taxonomies for lexical data collections are relevant for the development of all kinds of metadata registries.

6. References

CEN CWA 15045. (2004). Multilingual Catalogue Strategies for eCommerce and eBusiness, CEN/ISSS/WS/eCat.

DS 2394-1. (1998). Lexical data collections – Description of data categories and data structure – Part 1: Taxonomy for the classification of information types, Danish Standards.

Gómez-Pérez, Asunción; Mariano Fernández-López & Oscar Corcho. (2004). Ontological Engineering – with examples from the areas of Knowledge Management, e-Commerce and the Semantic Web. London: Springer Verlag.

Gruber, T.R. (1993) "A translation Approach to Portable Ontology Specifications". *Knowledge Acquisition*, 5(2), pp. 199-220.

Guarino, Nicola (1998). "Formal Ontology and Information Systems". I: N. Guarino (ed.): *Formal Ontology in Information Systems*. IOS Press, Amsterdam, pp. 3-15.

Ide, Nancy and Laurent Romary. (2004). A Registry of Standard Data Categories for Linguistic Annotation. In: *Proceedings of the Fourth Language Resources and Evaluation Conference* (LREC), Lisbon, pp. 135-39.

ISO 704:2000. Terminology work – Principles and methods. International Standards Organisation.

ISO IEC 11179-3. (2003). Information technology – Metadata registries (MDR) – Part 3: Registry metamodel and basic attributes.

ISO TC 37/SC 3 N542. (2005). New title and scope & Proposals for new projects.

Madsen, Bodil Nistrup, Hanne Erdman Thomsen & Carl Vikner (2004). 'Principles of a system for terminological concept modelling'. In : Lino, Maria Teresa; Maria Fransisca Xavier, Fátima Ferreira, Rute Costa, Raquel Silva (eds.): Proceedings of the Fourth International Conference on Language Resources and Evaluation. ELRA, pp. 15-19.

METeOR http://meteor.aihw.gov.au/content/index.phtml/itemId/181162 (last visited on April 1st 2008).

Ruiz, Francisco and José R. Hilera (2006) "Using Ontologies in Software Engineering and Technology". In: Calero, Coral; Francisco Ruiz & Mario Piattini (eds.). (2006). Ontologies for Software Engineering and Software Technology. Berlin Heidelberg: Springer Verlag, pp. 49-102.

Wright, Sue Ellen. (2004). A Global Data Category Registry for Interoperable Language Resources. In: *Proceedings of the Fourth Language Resources and Evaluation Conference* (LREC), Lisbon.

7. Appendix: Overview of DS 2394-1: Lexical data collections

This appendix includes an overview of the taxonomy for the classification of information types in lexical data collections developed by the Danish Standard.

Main group	Category	Subcategory
Administrative information	• Internal reference	
	• External reference	• Literature reference • Source reference
	• Information on the collection and processing of data (data management) • Technical information	
Etymological information	• Origin • Parallel	
Grammatical information	• Part of speech • Gender	
	• Information on inflection	• Stem • Paradigm information • Inflected form
	• Word formation	
	• Syntax	• Syntactic frame (valency) • Specification of syntactic frame • Specification of auxiliary verb • Syntactic function
Graphical information	• Orthographical information •	• Spelling • Hyphenation
	• Graphic symbol	
Language		
Phonetic information	• Prosodic features • Segmental features	

Main group	Category	Subcategory
Semantic information	• Subject classification	• Classification system • Normative subject classification • Nonnormative subject classification
	• Semantic relations	• Concept system • Position of concept in concept system • Generic relation • Partitive relation • Successive relation • Causal relation • Associative relation • Antonymy • Metonymy • Equivalence within one language • Equivalence between two or more languages • Equivalence constraint
	• Content specification	• Lexical paraphrase • Analytic definition • Denotative definition • Ostensive definition • Additional definition • Additional information • Background information • Factual explanation • Characteristic feature • Figurative meaning
Structural information	• External structure • Internal structure	
Usage	• Examples of usage	• Citation • Collocation
	• Information on usage	• Temporal • Spatial • Communicative • Frequency
	• Evaluative information	