

## From Thesauri towards Ontologies?

**Abstract:** The ISO 2788 guidelines for monolingual thesauri contain a differentiation of “the hierarchical relationship” into “generic”, “partitive”, and “instance”, which in view of the purposes of document retrieval was in principle allowed to be neglected or blurred. However, ontologies, designed as language inventories for a wider scope of knowledge representation, are based on all these and some more logical differentiations. Rereading the ISO 2788 standard and inspecting the published Cyc Upper Ontology, it is argued that the adoption of the document-retrieval definition of subsumption generally prevents the conception or use of a thesaurus as a substructure of an ontology of the new kind as constructed for AI applications. When a thesaurus is used for fact description and inference on fact descriptions, the instance-of relationship too should be reconsidered: It may also link concepts and metaconcepts, and then its distinction from subsumption is needed. The treatment of the instance-of relationship in thesauri, the Cyc Upper Ontology, and WordNet is described from this perspective.

### 1 Introduction

Well aware of I. Dahlberg’s philippic against the use of the term ‘ontology’ by computer scientists and against its decorative use by others<sup>2</sup>, we start with a citation from J. F. Sowa’s invited talk on “Ontologies for Knowledge Sharing”, given to the Vienna congress *Terminology and Knowledge Engineering* in 1996. Sowa presented the following informal explications of the terms ‘ontology’, ‘terminological ontology’, and ‘axiomatized ontology’ (Sowa 1996, 14):

- “*Ontology*: A specification of the kinds of entities that exist or may exist in some domain or subject area. Formally, an ontology is specified by a collection of names for concept and relation types organized in a partial ordering by the type-subtype relation. Formal axioms and definitions may be associated with some or all concepts and relations in an ontology.”
- “*Terminological ontology*: An ontology whose concepts and relations are not fully specified by axioms and definitions that determine the necessary and sufficient conditions of their use. The concepts may be partially specified by relations such as subtype-supertype or part-whole, which determine the relative positions of the concepts with respect to one another, but do not completely define them.”
- “*Axiomatized Ontology*: A terminological ontology whose concepts and relations have associated axioms and definitions that are stated in logic or in some computer-oriented language that can be automatically translated to logic. There is no restriction on the complexity of the logic that may be used to state the axioms and definitions.”

In view of the fact that there was a lengthy discussion of these terms in the ontology community, we do not want to embark on reviewing the various attempts to define them, instead of this, we are interested in describing and understanding a few basic phenomena associated with terms already in use. Sowa addressed WordNet and Cyc as large ontologies, and we may assume that he would classify WordNet as a terminological and Cyc as an axiomatized ontology. In this paper, we will use both as examples, although only the published small fraction of Cyc, known as the *Cyc Upper Ontology*, was accessible to us<sup>3</sup>.

For study purposes, we imported the Cyc Upper Ontology into our system TerminologyFramework. This system was designed as a generic framework for representing all sorts of thesaural structures and their operational semantics. It has been the host of the

following different thesauri: Art and Architecture Thesaurus (AAT), German Parliamentary Thesaurus – and WordNet 1.5 (see Fischer et al. 1996). Because in turn it is hosted by a frame-based representation system, it can also cope with unconventional structures. The currently available assertions or axioms of the published Cyc Upper Ontology could all be represented without loss of information in our system by links between frames, although some links expressed with a negation. However, we do not know what has been omitted even with respect to the published small subset. The definitions for concepts are given by comments in natural language, and the currently *available* assertions actually result in partial concept definitions only – as is true with a thesaurus! So far, the published part according to our view is not a full-fledged, but a clipped axiomatized ontology; however, there is sufficient material to grasp its spirit whether this may strike you as odd or as intriguing.

Sowa adds to his explications (Sowa 1996, 14): “Although a terminological ontology may be expressed in logic, the versions of logic required are usually simpler, less expressive, and more easily computable than full first-order predicate calculus. ... The distinction between terminological and axiomatized ontologies is one of degree rather than kind. Axiomatized ontologies tend to be smaller than terminological ontologies, but their axioms and definitions can support more complex inferences and computations.”

These words bring to mind not only the idea of relatedness of systems by formal containment or partial compatibility, but also the essential differences with respect to purpose and field of application: Thesauri are language inventories to be used for simple factual statements, often about documents and in principle of the one kind: “Document d (an individual of the domain) treats topic c (a concept, a language element of the inventory).” The thesaural synonymy and hierarchy serve as an access structure to documents and to broaden or narrow a query, which may be seen as a kind of inference.<sup>4</sup> On the other hand, axiomatic ontologies were designed as language inventories which allow expression of a much greater variety of facts in a greater variety of domains, and which also permit the control of their consistency and the inference of implications automatically. Accordingly, verb concepts or relationships, defined on the entity or the concept domain<sup>5</sup>, play a central role in ontologies, quite in contrast to thesauri.

However, if a thesaurus, at least with its essential hierarchical structure of noun concepts, can, in principle, be conceived as a kind of simple substructure of an ontology, are the other concepts, axioms and machine-readable definitions just *additional* information elements provided for the additional purposes? Or in practical terms, *under what conditions does it make sense to use the noun concept hierarchies of a thesaurus as a starting-point for an upgrade to at least a terminological ontology?* This is the constructive meaning of the paper’s evocative title question.

A main, however simple, argument in this paper is that the treatment of the hierarchy relationship (cf. Sowa’s uncommented reference to the (sub)type-supertype relationship) is the crucial point where opinions may differ and in practice may separate the world of thesauri and logic-oriented ontologies. This has also been a more or less hidden source of controversy among thesaurus practitioners, and not only with respect to this discussion we suggest that a study of the logic-oriented ontologies, such as the Cyc Upper Ontology, may be helpful and rewarding, whatever the conclusions.

## 2 “The Hierarchical Relationship” ?

In the standard specifications (ISO 2788, DIN 1463) as well as in the respective literature (Soergel 1974, Wersig 1978, Lancaster 1986) we find that the thesaural relations of the kind “hierarchy” are differentiated into “generic”, “partitive”, and “instance” (ISO 2788 and Lancaster only). However, the standards more or less implicitly allow that these different types of hierarchy relations may be conflated into one hierarchical relationship in an actual thesaurus; we see this also reflected in the title “The Hierarchical Relationship” (ISO 2788, 8.3).

## 2.1 The Document-Retrieval Definition of Subsumption

The rationale for this permissiveness lies in what we may call the document-retrieval definition of ‘broader-narrower’, as given by Soergel (1974, 78):

“Concept A is broader than concept B whenever the following holds: in any inclusive search for A all items dealing with B should be found. Conversely B is narrower than A.”

Soergel is well aware of the fact that this definition introduces subjectivity and consequently suggests that concrete hierarchical links are backed up by a majority count based on expert judgements or an analysis of search requests. He adds: “It is in order to emphasize the purely pragmatic nature of the definition. It is oriented toward the function of hierarchical relationships in the search process. This is not to deny that logical or philosophical considerations might be helpful in suggesting hierarchical relationships.” (Soergel 1974, 79)

An example for this is (hagiography BT saints). Soergel (1990, 14) found this link in an early version of the Art and Architecture Thesaurus and commented it in a note: “Purists would not use BT ... *saints* because *hagiography* is not a kind of saint. Pragmatically, the question is simply: Does a user searching under *saints* generally want to find documents on *hagiography*. If so, BT should be used. Alternatively one might use RT, but there should be a cross-reference.” In the 1991 version and in later versions of the AAT the mentioned link however was erased and replaced by (hagiography BT biography) in accordance with the AAT guidelines (1994) which refer to the genus-species or class-subclass relationship as the basis of the AAT hierarchies.

## 2.2 The Extensional Definition of Subsumption

The “logical considerations”, mentioned by Soergel, come into play, when other labels for ‘broader-narrower’ such as “genus-species” or “is kind of” (for ‘broader’) (Lancaster 1986) are used to characterize the generic hierarchy relation. ISO 2788 (8.3.4.1) recommends an “all-and-some”-test for this relation and explains it, using the examples BIRDS versus PARROTS and PARROTS versus PETS. Informally, it is said there that concepts are taken as classes which have members, and that for a genuine narrower concept all its members must also be members of the broader concept while for the broader concept only some of its members must also be members of the narrower concept. Accordingly, PARROTS is a subconcept of BIRDS, but not of PETS.

In the “all-and-some”-test we recognize the *extensional definition of subsumption*: A noun concept  $c$  is a unary predicate  $c$  and its extension  $\text{ext}(c)$  is the set of objects  $x$  for which  $c(x)$  is true. To be more precise we could add a time and/or context parameter such that we could speak of  $\text{ext}(c,t)$ , the set of objects for which  $c(x)$  is true at time / in context  $t$ . A concept  $c$  is narrower at time / in context  $t$  than concept  $b$  if  $\text{ext}(c,t)$  is a subset of  $\text{ext}(b,t)$ . In Cyc every assertion, including a subsumption statement, is bound to a context (“microtheory”) which can be organized in an inheritance hierarchy; in addition, Cyc allows expression of defaults and exceptions from defaults. Accordingly, we could express that in a special context all parrots are pets (cf. ISO 2788, 8.3.4.1), and we could express whether exceptions are allowed or not. This is a way to go beyond merely accepting that every thesaurus constitutes its own microtheory; it is a way to control idiosyncrasy and to support default reasoning.

In the comments, attached to unary Cyc predicates corresponding to nouns, very often the set/class model is used for explanation, even equating a concept and its extension. A typical example is the comment for the concept labeled #Bird: “The collection of birds; a subset of #Vertebrate. Each element of #Bird is an ... animal ...”<sup>6</sup>.

## 2.3 The Intensional Definition of Subsumption

In the German standard (DIN 1463) and the already mentioned German literature we find another formulation of generic subsumption which is based on the representation of concepts as

sets of property or attribute values (“Merkmale”, features): The narrower concept contains all the attribute values of the broader concept plus at least one in addition.

In this formulation we recognize a form of the *intensional definition of subsumption*. The concept-defining language may be generalized from attribute value sets to more complex expressions in a formal language which includes a syntactic inference mechanism for these expressions (see Woods, 1991). The subsumption relation then is defined inductively: The induction may start from axiomatic subsumption statements for undefined or partially defined concepts and continues by syntactic or structural inference on the concept defining expressions. In such an environment the inference is exerted by a program (the “classifier”), i.e. a subsumption statement for a fully defined concept is no longer at the knowledge engineer’s disposal. This is the world of KL-ONE-type systems or the field of terminological or concept logic (e.g. see Brachman et al., 1991). In terminological logic, both the intensional and extensional definition of subsumption are complementary as the syntactical formalism and its interpretation by a mathematical set model, i.e. the “meaning” of a an intensional subsumption can be found in the complementary extensional subsumption<sup>7</sup>.

#### 2.4 Interim Summary and Argument

We can take it for granted that the axiomatized ontologies are based on the extensional and/or the intensional definition of subsumption<sup>8</sup>. It follows from this that the generic hierarchy of a thesaurus is incompatible with that of an ontology *to the extent* that the thesaurus builders did not make use of logical subsumption, but deviated from it under the influence of the “floating” pragmatics of the document-retrieval definition of subsumption. We are facing the following two questions:

- How can we find out how many of these deviations are included in a given thesaurus?
- What is gained by the logical "purism"?

The first question will be dealt with in the next subsection; however, the answer cannot be satisfactory. For the second question we only give the following outline of an answer: Even if we would confine ourselves to classical document retrieval, our strategic argument in a few words is: The document-retrieval definition of subsumption results in a coarse mechanism to broaden and narrow a question. This has of course its counterpart in a coarse document representation, i.e. indexing technique. When ontologies of the new kind are combined with a concept-oriented lexical database, they may cover a spectrum of functionality which in principle includes all the traditional services of a classical thesaurus, and can offer more.

#### 2.5 Some Remarks on Checking Consistency of Subsumption in a Thesaurus

The individual thesaurus guidelines give us a clue what kind of subsumption the thesaurus builders adhered to, but that is not sufficient. In a conventional thesaurus, concept definitions which a computer program can “understand” are not available. “Understanding” would imply that the computer program from the definition, supplied by the knowledge engineer, would be able to “classify” the given concept, i.e. to link it in a non-redundant way to its superconcepts and subconcepts which already exist in the thesaurus – or have to be created as undefined, but presupposed by the given definition. In a conventional thesaurus, subsumption statements or links are given by the knowledge engineer, and these statements are only subject to formal consistency checks which result in *necessary, but not sufficient* conditions for correctness.

If only the generic hierarchy relationship is provided in a thesaurus, then the only classical formal criterion of consistency of this relationship is acyclicity. If further relationships of the hierarchical or associative type are provided, then further consistency rules can be formulated and used for checking (cf. Fischer 1993). How different conceptual relationships can hold each other in check, we exemplified for WordNet 1.5 (see Fischer, 1997). With respect to subsumption and antosemy we formulated two rules: 1. Subsumption and antosemy exclude each other. 2. Antosemic concepts must not have common subconcepts. These rules can be

justified by a feature model of concepts (see above 2.3). Checking the WordNet database as to these constraints, only one constellation contradicting rule 1 and three constellations contradicting rule 2 were detected. One counter-example of rule 2 was the verb concept ‘smuggle’ (to export or import illegally) as a troponym of (subconcept of) the verb concept ‘export’ and a troponym of the verb concept ‘import’. The superconcepts ‘import’ and ‘export’ are stated to be antosems (indirectly via antonymy between their terms or synset elements). Then the elements of the extension of ‘smuggle’, i.e. the smuggle events, must all be illegal export events and at the same time illegal import events which by default may be true, but need not be true<sup>9</sup>.

An analysis of the constellations contradicting rule 2 led us to the pattern of “disjunctive hypernyms”. We give the following short abstract description: If it is stated that a concept *c* has several superconcepts (*b*<sub>1</sub>, *b*<sub>2</sub>, ...), then the generic hierarchy requires that the “all-and-some”-test must be fulfilled for *each* superconcept. If only: “For all *x*: If *c*(*x*), then either *b*<sub>1</sub>(*x*) or *b*<sub>2</sub>(*x*) or ...”, then the *b*<sub>1</sub>, *b*<sub>2</sub>, ... could be called “disjunctive hypernyms”, but such a constellation needs either a special relationship or *one* superconcept which is the disjunction of the *b*<sub>1</sub>, *b*<sub>2</sub>, ..., if the extension of this one disjunctive predicate (*b*<sub>1</sub> or *b*<sub>2</sub> or...) corresponds to a covering of the extension of *c*. Otherwise the transitivity of generic subsumption is no longer valid.

We encountered the awareness of the constellation of “disjunctive hypernyms” in the thesaurus literature, but it was not marked as to be considered harmful: See Wersig (1978, 129), citing Heinzmann’s term “mehrzielig-mehrdeutige generische Beziehung” (multiple-target / ambiguous generic relation).

An important feature to control the consistency of fact descriptions or to make positive inferences, if, for example, a property ascription is negated, is the mathematical notion of covering or (stronger) partition of a set. The Cyc Upper Ontology includes predicates to express that a set of concepts is a covering or even a partition of a concept. This was obviously not needed for indexing and document retrieval because a document may be relevant with respect to different topics which, when property ascriptions for a single individual, may exclude each other. The ISO 2788 guidelines mention the possibility of checks based on disjoint concept hierarchies (see section 8.3.2), and the AAT vocabulary is grouped into disjunctive hierarchies according to facets, however, the AAT is based on generic monohierarchy. We are not aware of thesaurus applications where disjointness and covering information is really used for consistency control and inference.

### 3 The Instance Relationship

Now let us assume that we have got a thesaurus in which neither partitive relationships and the generic relationship are merged nor a document-retrieval type subsumption link was intended. Then we may still face a problem, if we want to use this thesaurus for inferences on fact descriptions, and this has to do with “the instance relationship”.

#### 3.1 The Instance Relationship as Treated by ISO 2788

According to ISO 2788 (8.3.6) the instance relationship “identifies the link between a general category of things or events, expressed by a common noun, and an individual instance of that category, the instance then forming a class-of-one which is represented by a proper name.” An example is given: ALPS and HIMALAYAS are instances of the class or concept MOUNTAIN REGIONS.

First we note that the ISO standard does not provide a special distinguishing label for this relation “instance-of” and its inverse, and this may be explained by the circumstance that “proper names are frequently excluded from thesauri, on the grounds that, if admitted, they would overload the categories” (ISO 2788, 8.3.6). In WordNet there are examples of instance-of relationships of this kind, and they are expressed by generic subsumption: (poet *hyponym*:

William Wordsworth) or inversely: (William Wordsworth *hypernym*: poet). The makeshift solution to use hypernym links for instance-of links may not be considered harmful in isolation as long as the system does not need to know the differentiation between a collection and an individual, and that the generic hierarchy should not be prolonged downwards beyond such an individual concept or individual<sup>10</sup>; however, we will show in the following by examples how this may be a source of fallacies or problems, when instance relationship and subsumption are merged and the structure is used for inference.

### 3.2 The Instance or Element Relationship as a Consequence of the Class Model of Concepts

In thesauri we do not only find a substitution of an instance-of link by a broader-link between individuals and concepts, but also between “genuine” or generic concepts. This may cause a fallacy, if we use such a hierarchy for logical inferences: Let us give an example from the German Parliamentary Thesaurus: (roofer *broader*: trade requiring an apprenticeship), in German (Dachdecker *weiter*: Ausbildungsberuf). All German trades which require an apprenticeship have been listed as narrower concepts of ‘trade requiring an apprenticeship’. However, for this relationship the “all-and-some”-test is not valid or adequate, if ‘roofer’ here is the same concept as in (Henry Smith *instance-of* roofer). The absurdity is obvious, if from the premises we correctly infer: (Henry Smith *instance-of* trade requiring an apprenticeship). Without taking resort to different meanings of ‘roofer’, this problem can be solved by taking seriously the class model of concepts and the “all-and-some”-test mentioned in ISO 2788: Then (roofer *broader*: trade requiring an apprenticeship) is not correct, but (roofer *instance-of* trade requiring an apprenticeship) is a solution, if we acknowledge that it makes sense to form classes of classes. This means: ‘trade requiring an apprenticeship’ is a concept whose extension is a set of concepts, and not a subset of persons. The concept ‘building craftsman’ could be an appropriate hypernym of ‘roofer’.

This formation of metaconcepts or concept containers (concepts which group concepts) is such a common mental activity that the absence and disregard of this kind of instance-of relationship in the standard is surprising. We do not see a reason, why we should not use the instance-of relationship both for individual/concept relations and for concept/ metaconcepts relations, and we are encouraged to do so, when we examine the use of the instance-of relationship in the Cyc Upper Ontology.

### 3.3 The Instance-of and Subclass-of Relationship in the Cyc Upper Ontology

Looking at the Cyc Upper Ontology, we notice that generic hierarchy and instance relationship are differentiated: A broader relation is labeled *subclass-of* (OKBC format) or *genls* (Cyc format), and an *instance-of* relation has just this label (OKBC format) or is labeled *isa* (Cyc format). Furthermore, the instance relationship is heavily used: There are 4,193 non-redundant instance-of links and 2,311 non-redundant subclass-of links (the ratio is 6,351 : 3,041, if we include instance-of or subclass-of type links for arguments of predicates and functions). Note that nearly all instance-of links in the Cyc Upper Ontology are *not* of the kind (ALPS instance-of MOUNTAIN REGIONS), but of the concept- metaconcept type; currently the file contains only 15 instances of the concept ‘entity’ (which represent in Cyc the “individuals” in the sense of ISO 2788), and this merely for the purpose of illustration.

As an example of an instance-of relation in the Cyc Upper Ontology, we take up the Cyc concept `#$Bird` (see Figure 1): The comment for `#$Bird` as well as the formal assertions state that `#$Bird` is an instance of `#$BiologicalClass`, furthermore, `#$BiologicalClass` is an instance of `#$BiologicalTaxonType`, and the co-elements of `#$BiologicalClass` in this collection are `...Division`, `..Family`, `...Genus`, `...Kingdom`, `...Order`, `...Phylum`, `...Species`, `...Subclass`, `...Subkingdom`. On the other hand, `#$BiologicalClass` is a subclass of `#$BiologicalTaxon`, and a common superclass of `#$BiologicalTaxon` and `#$BiologicalTaxonType` is `#$ConventionalClassificationType`. So we can say: The concept `#$Bird` and the concept

`#$BiologicalClass` are both instances of `#$ConventionalClassificationType`, but the concept `#$Bird` is not a `#$BiologicalTaxonType`; positively, `#$Bird` is a `#$BiologicalTaxon`<sup>11</sup>. The logic of this inference is treated in the next subsection.

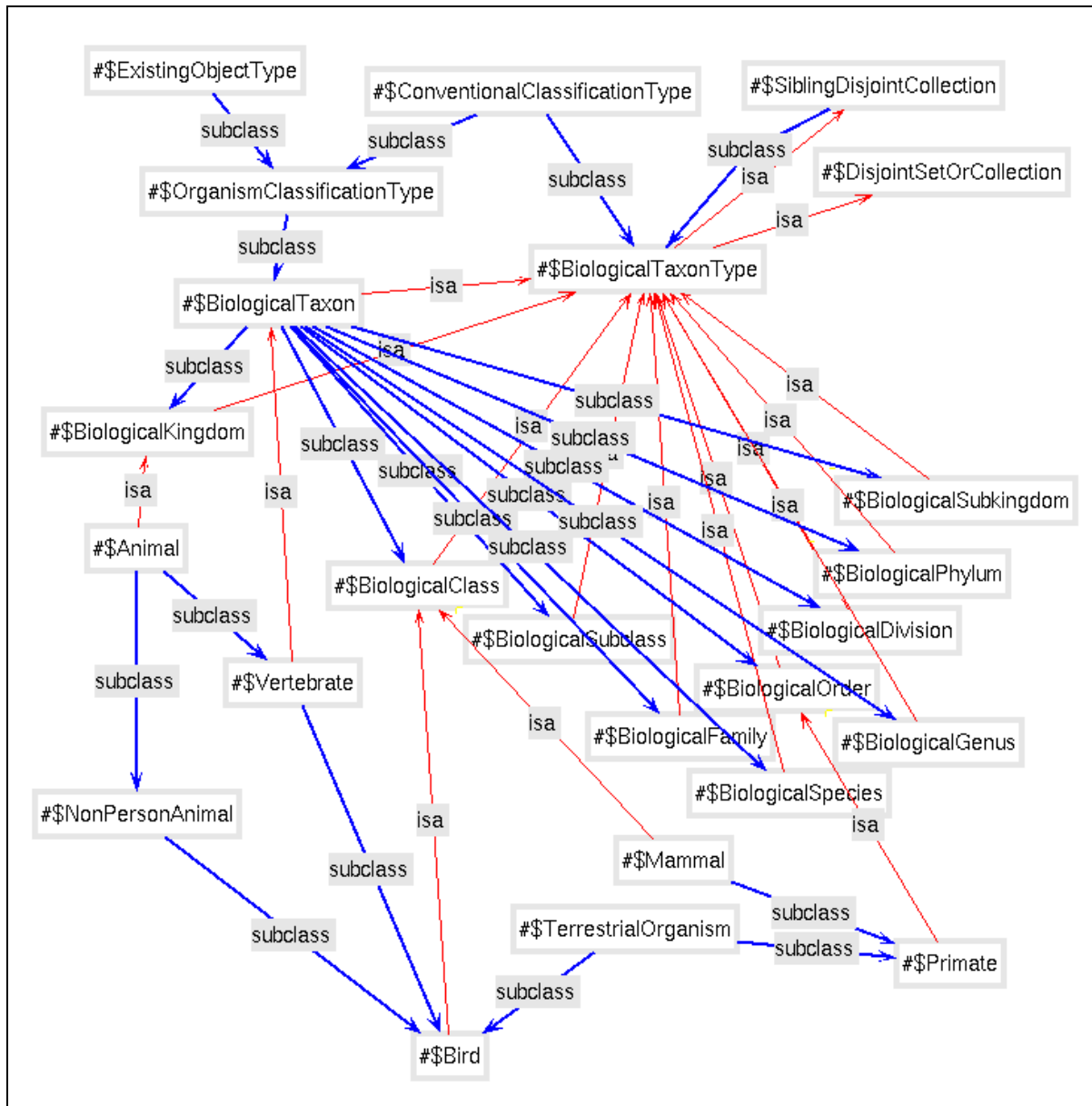


Figure 1: Examples of the instance-of (*isa*) relationship in the Cyc Upper Ontology

### 3.4 Some Formal Properties of the Instance-of and Subclass-of Relationship

According to Cyc, the instance-of relationship is a specialization of the mathematical *element-of(-a-set)* relationship, and the subclass-of relationship is a specialization of the mathematical *subset-of(-a-set)* relationship.

From the mathematical set model we can derive that the subclass-of relationship has the formal property of transitivity while this is not true for instance-of. If we join both relationships, transitivity is invalid for the merged relationship. The validity of transitivity for the subclass-of relationship enables inheritance and transmitting of data along the generic hierarchy. Assertions valid for all elements of the extension of a concept can be transmitted down the subset/subclass hierarchy, but an assertion valid for an element itself – which itself is a set – is not necessarily valid for the elements of this set.

Although transitivity for the instance-of relationship does not hold, an instance-of link is transmitted upward via the subclass-of links: I.e., if (A instance-of B) and (B subclass-Of C) directly or indirectly, then also (A instance-of C); for example, if Henry Smith is a roofer (instance-of link), and every roofer is a building craftsman (subclass-of link), then Henry Smith is a (instance-of link) building craftsman.

A chain of instance-Of relations could be shown as a hierarchy, but that would be deceptive, if we assume transitive connections as we do for the generic hierarchy. There are in fact instance-of chains in the Cyc Upper Ontology actually up to a length of 5, and a new subclass / subclass-of chain (maximal length 16) can start from each metaconcept. From the scores given for the Cyc Upper Ontology we may infer that the instance-of relationship is much more relevant for a logic-based knowledge representation than we would expect from their neglect in the thesaurus guidelines. The metaconcepts often correspond to technical, artificial or non-lexicalized terms which are introduced to factorize information for their elements, needed in applications we can only guess.

The class or set model of concepts does not exclude that both an instance-of link and a subclass-of link may connect two concepts. In fact we find 11 parallel link pairs of this kind in the Cyc Upper Ontology. They all connect metaconcepts: One example is #BiologicalTaxonType (see Figure 2) which is both an instance of #SiblingDisjointCollection and a subclass of #SiblingDisjointCollection. We do not present the formal definition of #SiblingDisjointCollection, but as the name partially indicates, it is a collection of collections which are disjoint in pairs – except for instances which are also instances of a common subclass. According to the given assertions, not only #BiologicalTaxonType is a #SiblingDisjointCollection, but this is true also for the elements of #BiologicalTaxonType.

### 3.5 The Instance-of Relationship in WordNet

In order to find out how WordNet treats concept-metaconcept relations and to compare this with Cyc, we examined WordNet's 'bird' concept field; an excerpt is presented in Figure 2<sup>12</sup>. With this picture in view, our first hypothesis was that WordNet uses the member-of relationship to express this type of instance-of relations. However, a more detailed analysis led us to discard this assumption.

On the left side of the graph we see that member-of links are used to link "familiar" biological concepts to scientific biological grouping concepts, for example 'bird[1]' member-of 'Aves', or 'European goatsucker' (WordNet synonym: 'Caprimulgus europaeus') member-of 'Caprimulgus'. We make the following objections:

- We see a questionable doubling of conceptual hierarchies. According to our view, the extension and intension of 'bird[1]' and 'Aves' (or 'caprimulgiform bird' and 'Caprimulgiformes') may logically be treated as identical. WordNet's gloss for 'Aves' is monosyllabic: "birds". In WordNet the biological class 'Aves' contains as members the scientific animal orders 'Anseriformes', 'Caprimulgiformes' etc. (not shown), and as a co-member (!) the "familiar" biological concept 'bird[1]' which in fact corresponds to a biological class (see Cyc). An analogous constellation we have for example for 'Caprimulgiformes': It contains the bird families 'Caprimulgidae', 'Podargidae', 'Steatornithidae' (not shown), and (!) the concept 'caprimulgiform bird' as members. We do not see a justification for treating the pairs 'bird[1]' - 'Aves', and 'caprimulgiform bird' - 'Caprimulgiformes' (as many others of this type) as member and member containing group.
- Between the nodes of the middle column from 'Caprimulgus' up to 'Animalia' and the nodes of the right column from 'bird genus' up to 'kingdom [2]' we see hyponym links where instance links would be adequate, for example between 'Aves' and 'class[5]' (biological class). The WordNet gloss for 'kingdom[2]' is: "one of five biological categories: Monera or Protista or Plantae or Fungi or Animalia". In Figure 2 we have shown only 'Animalia' as one of the (in fact) eight<sup>13</sup> existing *hyponyms* of 'kingdom [2]'. This is

additional evidence that the hyponym link in WordNet is also used to express an instance link between metaconcept and concept.

- If in this constellation we sensibly would merge, for example, the concepts ‘bird[1]’ and ‘Aves’, or ‘caprimulgiform bird’ and ‘Caprimulgiformes’, we would get fallacies: For example, we could infer that the concept ‘European goatsucker’ is a hyponym of ‘class[5]’ (biological class) because the inference cannot know that the hyponym link between ‘class[5]’ and ‘Aves’ is an instance link which interrupts transitivity. *Given that the instance and the hyponym relationship have not been differentiated in WordNet*, was the avoidance of these kinds of fallacies the cause of the enigmatic split of ‘bird[1]’ - ‘Aves’ or ‘caprimulgiform bird’ - ‘Caprimulgiformes’, and their linkage by member-of links?

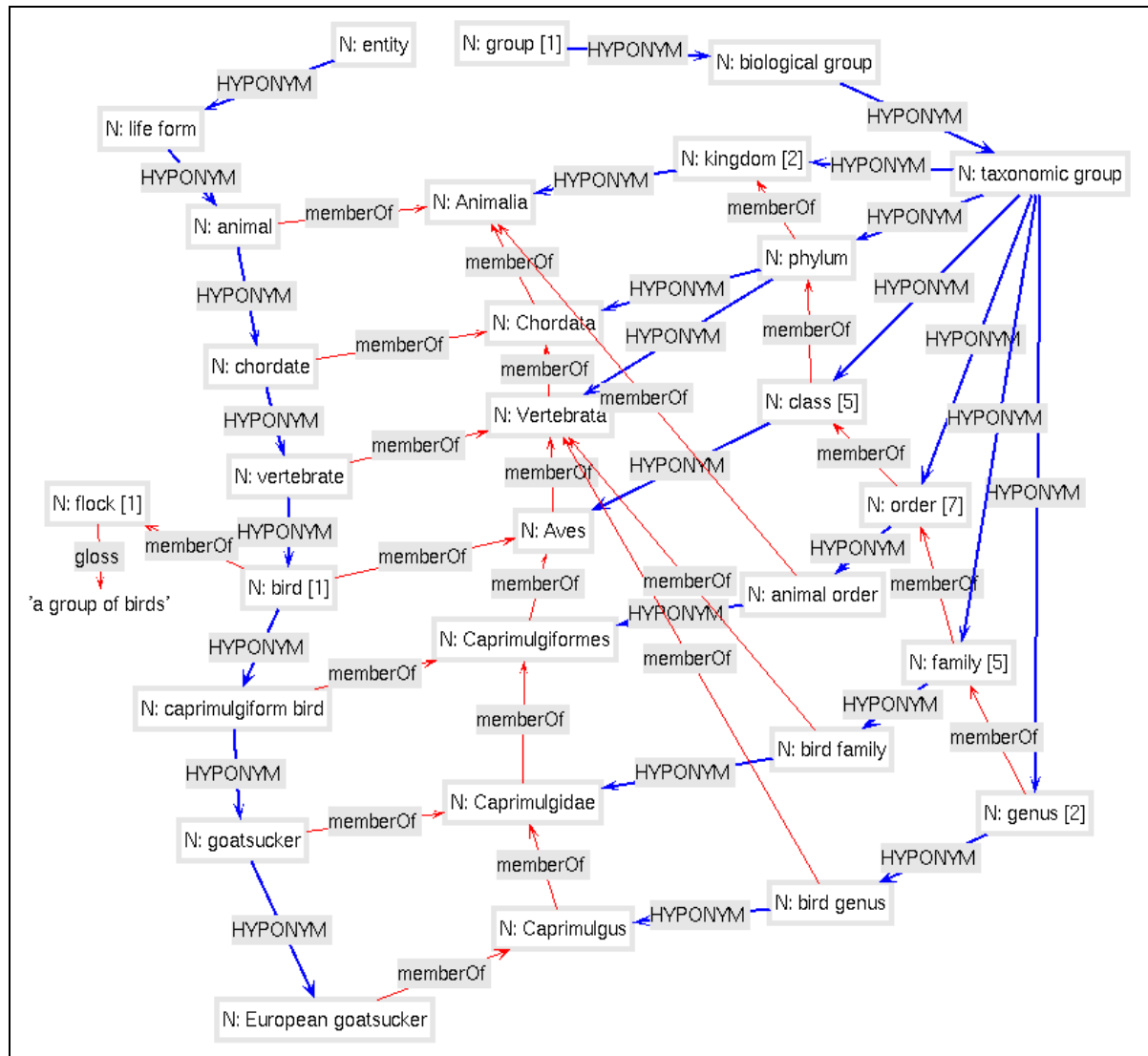


Figure 2: Excerpt from the concept field ‘bird’ in WordNet

In addition, WordNet wants to represent that each taxonomic family is a grouping of subfamilies or genera, and this is expressed by: (genus[2] *memberOf* family[5]) and (subfamily *memberOf* family[5]). On the one hand, we do not see that this kind of knowledge, although mentioned in the Cyc comments, actually is represented in the Cyc Upper Ontology. On the other hand, WordNet thereby introduces another meaning of a member-of link; the intended meaning may be expressed by: For all x: If x is a taxonomic genus, then exists a taxonomic family or a subfamily y such that x is a member of y, etc.. This is, for example, different from a member-of statement about a membership of an individual (abstract or concrete) in an

individual group (abstract or concrete), as we have it in: The bird genus *Caprimulgus* is a member of the bird family *Caprimulgidae*<sup>14</sup>. Other meanings, existing in WordNet, are: (Some!) birds are members of a (some!) flock (see Figure 2), or: (All?) Amish are members of the (individual!) Amish sect. Cyc provides the logic language to express these different forms of assertions adequately; if we use a “node-and-link” language we could say that the base link type ‘memberOf’ needs “quantificational tagging” for meaning differentiation (see Woods, 1991; Priß, 1996).

### 3.6 Thesaural Guide Terms and Node Labels May Be Concept Containers / Metaconcepts

The “guide terms” or facet node labels, used to group sibling concepts in a thesaurus (see ISO 2788, 8.3.3 and 9.3.3. a), often can be interpreted as metaconcepts or concept containers. We do not see a way how to describe ‘(CARS) by motive power’ as a subset of cars, and thus as a subconcept of CARS. On the contrary, it is adequate to paraphrase it by “The *collection of car collections* which you get, when you group cars by their motive power”; then, by our own words, we describe a *set of car sets*, i.e. a set of concepts which are subconcepts of CARS, i.e. we describe a metaconcept which for example contains the concepts DIESEL CARS and ELECTRIC CARS as elements.

Such a concept-container interpretation is quite clear, if the guide node is labeled ‘<concepts in the arts>’; however, the contrary may be true for ‘<administrative bodies>’, which might well be interpreted as a concept whose extension contains all administrative bodies; both are guide term examples from the Art and Architecture Thesaurus, the angled brackets identifying them as non-descriptors. On the other hand, if a language generation system would use that hierarchy then it should know, that it is legitimate to use the term ‘administrative body’ as a more general referential expression for an individual X which in the context of the discourse was formerly referenced as for example a ‘county’, while it would not be legitimate to address an X, characterized as a ‘satire’, also as a ‘concept in the arts’.

## 4 Conclusion and Outlook

This paper was an attempt – from the information science perspective – to contribute to an understanding of the relationship between classical thesaurus guidelines and practices, and phenomena we can now observe in ontologies of the new kind. There have been voices in the field of information science who strongly recommended basing a thesaural hierarchy on the logical definition of subsumption for the sake of the “demanded compatibility in cooperating information retrieval systems” (Rolland, 1973, 93). This has also been a point of unsettled controversy in the discussions of the German Committee for Thesaurus Research (KTF). For the purpose of classical document retrieval the differentiations of “the” hierarchy seemed to be irrelevant; however, they are relevant, if we want to promote a refinement of document retrieval or an alignment, merging or upgrade of thesauri or conceptual knowledge bases to be used and re-used for a wider spectrum of applications.

Of course, there are other relevant aspects we could not deal with here. For instance, observing that among the approximately 20.000 concepts of the 1995 version of the German Parliamentary Thesaurus 25% are top concepts with respect to BT<sup>15</sup>, then it would mean a great effort to remodel it into an ontology for which we expect a more pyramidal form of the conceptual hierarchy. Furthermore, at the example of WordNet’s member-of relationship we could only touch the topic of “the hierarchical whole-part relationship” (ISO 2788, 8.3.5). This topic needs an extensive separate treatment (see Priß 1996) as well as the thesaural approach to concept definitions as proposed by Rahmstorf (1994). Rahmstorf’s concept definition scheme, based on “pure” subsumption and a differentiation of the coarse thesaural appurtenance relationship by a given set of conceptual relationships, we see still more differentiated in the concept definition languages of the KL-ONE-type family of knowledge representation systems.

## References

- Brachman, R.J., Mc Guinness, D.L., Patel-Schneider, P.F., Resnick, L.A., Borgida, A. (1991). Living With Classic: When and How to Use a KL-ONE-like Language. In: Sowa, J.F.(Ed.): Principles of Semantic Networks. San Mateo, CA: Morgan Kaufmans Publ., 401-456.
- DIN 1463 Teil 1 (1987). Erstellung und Weiterentwicklung von Thesauri. Einsprachige Thesauri.
- Fischer, D.H. (1993). Consistency Rules and Triggers for Multilingual Terminology. In: Schmitz, K. D. , (Ed.): TKE'93: Terminology and Knowledge Engineering. Frankfurt/M: INDEKS Verl., 333--342.
- Fischer, D.H.; Möhr, W.; Rostek, L. (1996). A Modular, Object-oriented and Generic Approach for Building Terminology Maintenance Systems. In: Galinski, C. and Schmitz, K. D. (Eds.): TKE'96: Terminology and Knowledge Engineering. Frankfurt/M; INDEKS Verl., 245-258
- Fischer, D.H. (1997). Formal redundancy and consistency checking rules for the lexical database WordNet 1.5. In: VOSSSEN. P./ADRIAENS, G./CALZOLARI, N./SANFILIPPO, A./WILKS, Y. (Eds.): Automatic Information Extraction and Building of Lexical Semantic Resources for NLP Applications. ACL/EACL-97 Workshop Proceedings, July 12 1997, Madrid. acl@bellcore.com, 22-31.
- ISO 2788 (1996). Documentation – Guidelines for the establishment of monolingual thesauri. 2. edition, 1986-11-15, Reconfirmed 1996.
- Lancaster, F. W. (1985). Thesaurus Construction and Use. A Condensed Course. Paris: General Information Programme and INISIST, PGI-85/WS/11.
- Miller, G.A., Beckwith, R., Fellbaum, C., Gross, D., Miller, K., Teng, R. (1993). Five Papers on WordNet. CLS Report 43, Cognitive Science Laboratory, Princeton University, July 1990, Revised August 1993. Available at <http://www.cogsci.princeton.edu/~wn/>
- Petersen, T.; Barnett, P.J. (Eds.) (1994). Guide to Indexing and Cataloging with the Art & Architecture Thesaurus. New York / Oxford: Oxford Univ. Press
- Priß, Uta (1996). Relational Concept Analysis: Semantic Structures in Dictionaries and Lexical Databases. Thesis Technical Univ. Darmstadt, Germany. Available at: <http://php.indiana.edu/~upriss/research.html>
- Rahmstorf, G. (1994). A New Thesaurus Structure for Semantic Information Retrieval. In: Finding New Values and Uses of Information. Proc. of the 47th FID General Assembly, Tokyo, Oct. 6-8 1994, 114-121
- Rolland, Maria Theresia (1973). Thesaurusprobleme in Informationsverbundssystemen. Pullach bei München: Verlag Dokumentation
- Soergel, D. (1974). Indexing Languages and Thesauri: Construction and Maintenance. Los Angeles, CA: Melville Publ. Company
- Soergel, D. (1990?). The Arts and Architecture Thesaurus (AAT). A critical appraisal. Manuscript.
- Sowa, J. F. (1996). Ontologies for Knowledge Sharing. Manuscript of the invited talk at TKE '96, Vienna 1996
- Wersig, G. (1978). Thesaurus-Leitfaden. Eine Einführung in das Thesaurus-Prinzip in Theorie und Praxis. München / New York: Verlag Dokumentation Saur KG
- Woods, W.A. (1991). Understanding Subsumption and Taxonomy: A Framework in Progress. In: Sowa, J.F. (Ed.): Principles of Semantic Networks. San Mateo, CA: Morgan Kaufmans Publ., 45-94.

## Notes

- <sup>1</sup> I am indebted to Wiebke Möhr, John A. Bateman, and Renato Reinau for valuable help to improve this paper.
- <sup>2</sup> See the editorial to *Knowledge Organization*, Vol. 23, 1996, No. 3, 129.
- <sup>3</sup> The published part currently comprises a little less than 3,000 concepts and 10,000 links, and is obviously the main source of a "Reference Ontology", the "Draft Merged Upper Level" of the ANSI Ad Hoc Group on Ontology Standards. As the names indicate, this group's efforts are dedicated to standardize and exchange reusable parts of knowledge bases, and the ontology is dedicated to be used on top of different more specialized ontologies; currently it also contains some concepts from another ontology used in linguistics which are either asserted to be equivalent with Cyc concepts or an attempt is made to state a logical relationship to one or more Cyc concepts. The "Draft Merged Upper

Level” ontology can be downloaded via <http://www.kls.stanford.edu/onto/std/> in three different formats: Cyc file format, KIF format and OKBC format. The Cycorp License Agreement is: “Cycorp is providing this material from the Cyc(tm) Upper Ontology at no charge, for everyone to use, including commercial service use and incorporation into products. However, it is not 'Public Domain.' Please acknowledge Cycorp, 3721 Executive Center Dr., Austin, TX 78731 in any use or citation of this material, and request that each further user include a full copy of this notice as well, in any use or citation they make of the material. All these terms equally apply to renamings and other logically equivalent reformulations of the material in any natural or formal language. Cycorp intends to amend and expand the material from time to time; the latest version is available at <http://www.cyc.com>.”

- <sup>4</sup> Note that an ontology as a *language inventory* in general only gives the means to express facts and defines the conditions of what *possibly* can be said to be true. Therefore the fact database (in our analogy: the indexed document pool) would not belong to the ontology (the thesaurus and its rules of usage).
- <sup>5</sup> For example, the Cyc Upper Ontology does not only include and describe the binary predicate ‘mother’, but also the ‘subclass-of’ and ‘instance-of’ relationship themselves (see below) which are used for concept description.
- <sup>6</sup> The concept names in the Cyc Upper Ontology are unique formal identifiers.
- <sup>7</sup> Note that the duality allows for intensional different concepts which have the same extension (cf. the ‘morning star’ and ‘evening star’ example).
- <sup>8</sup> With some reservations this may also be stated for WordNet’s hyponymy relationship for noun senses (noun concepts) and troponymy relationship for verb senses, see Miller et al. (1993). Miller et al. (8) refer to an intensional, feature based model of hyponymy for nouns, while Fellbaum (47) refers to an extensional model for verb troponymy. However, Miller et al. (8) also base subsumption on the “a kind of” test: “A concept represented by the synset {x, x',...} is said to be a hyponym of the concept represented by the synset {y, y',...} if native speakers of English accept sentences constructed from such frames as *An x is a (kind of) y.*” The homonymy of this test blurs the differentiation of the subclass-of and instance-of relationship and may be a source of observed fallacies or distorted constellations (see below). Our cited analysis of WordNet (Fischer et al. 1996, Fischer 1997) pertained to release 1.5, but the examples, given in this paper, are apparently unchanged in release 1.6.
- <sup>9</sup> If one argues that ‘default’ is the (always?) intended truth value strength (cf. Cyc) of the subsumption link, we ask what then is the truth value strength of the antosemy link?
- <sup>10</sup> We do not embark here on a discussion of the terms ‘individual concept’ versus ‘individual’.
- <sup>11</sup> Note that Figure 1 is incomplete at most of the nodes; however, it is complete at node Bird, BiologicalTaxon and BiologicalTaxonType with respect to *isa*, *subclass*, and the respective inverse relationships.
- <sup>12</sup> Note that the picture is complete with respect to the shown relationships only if we read it from bottom to top. The node label prefix ‘N:’ indicates a noun concept.
- <sup>13</sup> Two of them, ‘Plantae [1]’ and ‘Plantae [2]’, obviously are genuine technical duplicates, not homographs or homonyms.
- <sup>14</sup> Cyc uses the binary predicate # $\$$ groupMembers to enumerate members of a group, however for Cyc every instance of # $\$$ Group is an # $\$$ Individual which cannot have an extension or a subclass, but will have members.
- <sup>15</sup> The number reduces to 7% if we include another idiosyncratic quasi-hierarchical relationship, a mixture of a part-of and appurtenance relationship.