Chapter 7: Classifications

Ludger Jansen

It has long been a standard practice for the natural sciences to classify things. Thus, it is no wonder that, for two and a half millennia, philosophers have been reflecting on classifications, from Plato and Aristotle to contemporary philosophy of science. Some of the results of these reflections will be presented in this chapter. I will start by discussing a parody of a classification, namely: the purportedly ancient Chinese classification of animals described by Jorge Luis Borges. I will show that many of the mistakes that account for the comic features of this parody appear in real-life scientific databases as well. As examples of the latter, I will discuss the terminology database of the National Cancer Institute (NCI) of the United States, the *NCI Thesaurus*.

1. *Chinese Animals, or How to Make a Good Taxonomy*

In a certain Chinese Encyclopedia, the *Celestial Emporium of Benevolent Knowledge*, as Jorge Luis Borges tells us (1981), the following taxonomy of animals can be found:

(1) those that belong to the Emperor
(2) embalmed animals
(3) trained animals
(4) suckling pigs
(5) mermaids
(6) fabulous animals
(7) stray dogs
(8) those animals included in the present classification
(9) animals that tremble as if they were mad
(10) innumerable animals
(11) animals drawn with a very fine camelhair brush
(12) others
(13) animals that have just broken a flower vase
(14) animals that from a long way off look like flies

This taxonomy is a sophisticated piece of literature. It is also a good example of a bad taxonomy. For the sake of brevity, I will call Borges's taxonomy 'CAT' for 'Chinese Animal Taxonomy'. What lessons can we learn from CAT? Here are some of the rules for good and useful taxonomies, which CAT contravenes:

- **Ontological Grounding:** Good taxonomies classify things on the basis of traits belonging to those things. This precludes meta-types such as type (12): *others*. Things do not belong to the *other* group because they have some particular trait (of being other). Similarly, (14) does not classify things on the basis of traits belonging to those things themselves, but on the basis of their appearance to an observer.
- **Structure:** Good taxonomies take into account the fact that types of things have subtypes: for example, in biology there are genera and species. In CAT, however, all types have equal standing. It could be argued that mermaids are fabulous animals, in which case (5) would need to be rendered as a subtype of (6).
- **Disjointness:** If we have such a hierarchy of types and subtypes, then anything that instantiates a subtype also instantiates the type of which it is a subtype. For example, in biological systematics, every animal that is a horse is also a mammal. However, types on the same level of biological classification should be disjoint: no animal is both a mammal and a reptile, or both a vertebrate and an invertebrate. CAT's types, however, do not meet this criterion: Type (1) animals that belong to the Emperor probably include trained animals belonging under heading (3) as well.
- **Exhaustiveness:** Good taxonomies subsume all the entities they purport to subsume. At times this can be difficult to achieve, as in the biological sciences where new species are often discovered in the course of empirical research. CAT, however, seems to be far from exhaustive, if we ignore the fact that we can put any animal whatsoever under heading (12), *others*. Sometimes, exhaustiveness and disjointness are grouped together as the *jointly exhaustive and pairwise disjoint* (JEPD) criterion of classification.
- **No ambiguity:** Good taxonomies do not use terms ambiguously. Fabulous animals, pictures of animals, and dead animals, however, are not animals, at least not in the same sense that pigs or dogs are animals. For this reason, the headings (2), (5), (6), and (11) do not fit into this schema. What is more, painted animals are not animals, but rather paintings in which animals are represented.
- **Uniformity:** Good taxonomies have a well-defined domain. The traits by which they classify their objects should be of a uniform kind and be exemplified throughout the domain. CAT, however, draws on

the distinguishing traits of several different kinds at once. Heading (1) sorts animals according to their owners, (4) according to species membership, among other things, (7) according to species membership plus the lack of an owner, (9) according to behavior, (13) according to the effects of behavior, and (14) according to an animal's appearance to a remote observer.

- **Explicitness and precision:** Good taxonomies are explicit and precise. Headings such as (12), others, fulfill neither criterion.
- **No meta-types:** Good taxonomies avoid meta-types that come about through the classification process itself. In CAT, heading (8) is such a meta-type, and any animal belonging to CAT belongs under heading (8). If all animals belong to CAT, then all animals belong under (8). Thus every animal that belongs under (8) also belongs under headings (1)-(7) or (9)-(14). If an animal belongs to CAT but does not feature under these headings, this is no problem at all. It can also belong to CAT if it is a member of heading (8) alone. Heading (8) is a very peculiar heading for a taxonomy.

Classifications containing such types as (8) lead to problems that correspond structurally to the semantic paradox engendered by the sentence (T): 'This sentence is true'. (T) is indeterminate with regard to its truth value (that is, it is neither determinately true nor determinately false) because every truth value will fit. If we assume that it is true, what it says is the case, i.e. that it is true, and that is what is required for it to be a true sentence. But if we assume that it is false, then, as with any false sentence, what it says is not the case. Each of the two truth-values, true and false, can consistently be attributed to (T).

In the same manner, whether or not we classify animals that do not belong to other CAT-types under (8) can only be determined arbitrarily. A good classification system should not allow for this kind of arbitrariness concerning which objects fit under its types. Things get worse with CAT*, which we might call a Russellian version of CAT, containing (8*) 'Animals that do not belong to CAT*' instead of (8). A type like (8*) leads to problems that correspond, structurally, to Russell's antinomy or the liar paradox: if an animal belongs to types (1)-(8) or (9)-(14), then it belongs to CAT* and thereby does not belong to (8*). This is clear. But if an animal does not belong to these types, we encounter a paradoxical situation. For if an animal did not belong to (8*) either, it would not belong to any CAT*-type at all, and so would not belong to CAT*. Animals that do not belong to CAT*, however, belong to (8*). If we suppose that the animal does not belong to the other types, it follows that, if something does not belong to (8*), then it belongs to (8*). But anything that belongs to (8*) belongs to CAT*. So the animal in question does *not* belong to (8*) after all. Classification systems should eschew such situations whenever possible.

2. *Medical Information Systems, or How to Make a Bad Taxonomy*

We have used CAT as a heuristic tool to point out some of the mistakes that can be made in the construction of a classification system. These mistakes appear, not only in literary parodies like CAT, but also in actual scientific practice. I will show this in the following, by discussing the *National Cancer Institute Thesaurus* (NCIT). This will provide the opportunity to discuss the abovementioned mistakes in greater depth, as well as to propose some ways of repairing them.

The National Cancer Institute in the United States created the NCIT to support its battle against cancer by developing an online controlled vocabulary for annotating and indexing information relevant to cancer research (Fragoso, *et al.*, 2004; see also Ceusters, Smith and Goldberg, 2005). It contains more than 110,000 expressions and 36,000 terms of importance to cancer research, including 10,000 types of medical findings and disorders, more than 5,000 anatomical kinds, upwards of 3,500 chemicals and medicines, and approximately 2,000 types of genes.

2.1. *Structuredness: Groups and Animals*

Whereas CAT is totally unstructured, the NCIT does have a hierarchy of supertypes and subtypes. Nevertheless, in many places the NCIT is unstructured, and it is sometimes structured incorrectly. Consider, for example, the NCIT entry ‘Subgroup’, which NCIT defines as a ‘subdivision of a larger group with members often exhibiting similar characteristics’. We should suppose that subgroups are groups, and this would indeed be implied by the NCIT definition of group, which is: ‘Any number of entities (members) considered as a unit’. But this link between ‘Subgroup’ and ‘Group’ – an important bit of structure – is missing from the NCIT.

This example, also, shows that the NCIT is sometimes structured incorrectly. For example, as the supertype of ‘Subgroup’ NCIT gives ‘Grouping’, which it defines as a ‘system for classifying things into groups or the activity of putting things together in groups'. But, as philosophical tradition knows (see for example Aristotle, *Categories* 3), the definition of the supertype must also be applicable to all its subtypes. Thus from the definition of 'Grouping', and from the fact that *Group* is considered to be a subtype of *Grouping*, we get the following absurd conclusion, that a *subgroup* is either a *system* for classifying things into groups or an *activity* of putting things together in groups.

The NCI *Thesaurus*'s classification of animals is of similar quality to Borges's CAT. In the NCIT, the type *animal* splits into the subtypes *invertebrate*, *laboratory animal*, *vertebrate*, and *poikilotherms.* The subtypes *vertebrate*/*invertebrate* already present a problem, since they are an exhaustive division of all animals (and a division frowned upon by some biologists). Second, the artificial type *laboratory animal* stands out inappropriately when listed alongside the three natural classes, since laboratory animals do not comprise a natural kind. The subdivision appeals to traits of a range of different sorts. Finally, in reality *poikilotherms* is a subtype of *vertebrate* and, so, should not be classified at the same level as its supertype.

2.2. *Disjunctiveness and Exhaustiveness: Patients*

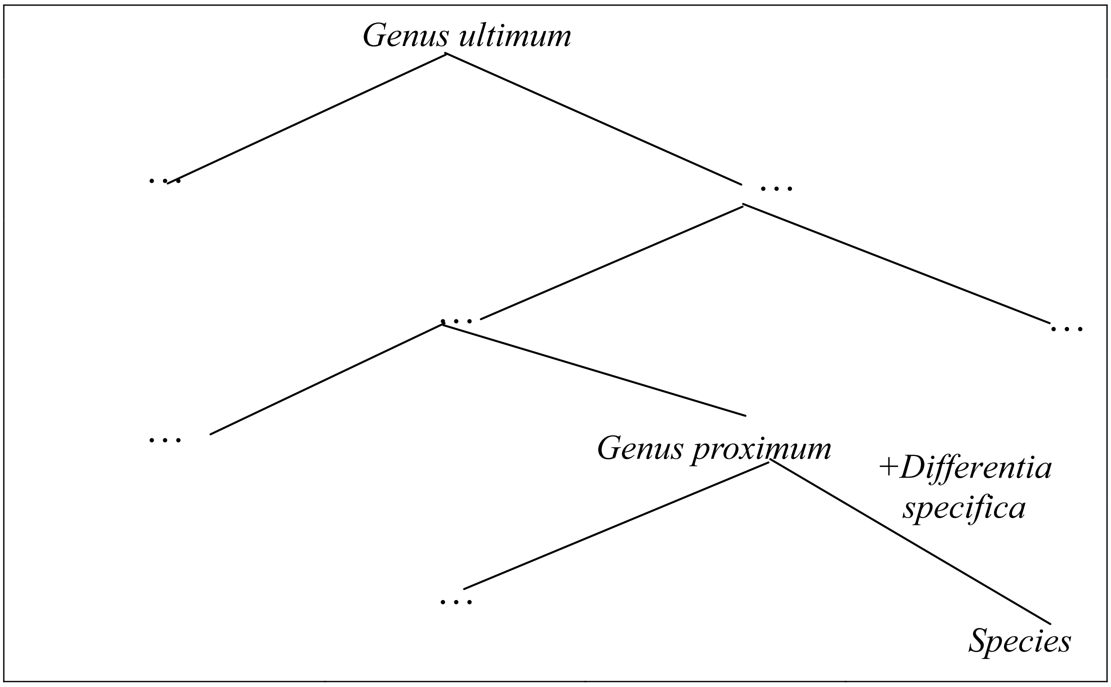
NCIT often contains subtypes which are not disjoint under the same supertype. An example is the entry *patient*. This entry has two subtypes: *cancer patient* and *outpatient*. These two entries are not disjoint, for many cancer patients are treated as outpatients. And naturally, these two subtypes are not an exhaustive classification of patients. There are many patients who are neither cancer patients nor outpatients. Normally, we would regard this example as a typical case of cross-classification, as there are two traits that an object could have independently of one another. Combined, these traits yield four classes of patients, as is presented in Figure 1:

Figure 1: *Four Classes of Patients: A Cross-classification*

PATIENTS	Outpatient? *Yes*.	Outpatient? *No*.
Cancer? *Yes*.	Outpatient with cancer	Inpatient with cancer
Cancer? *No*.	Outpatient without cancer	Inpatient without cancer

Classification systems are often constructed in such a way as to have the structure of an inverted tree, with a single highest-level root node and all nodes beneath this root having at most one single parent node. This practice derives from the long tradition of the Porphyrian tree, named after the neo-Platonist Porphyry (ca. 234–304), whose introductory guide to Aristotle's *Organon*, the *Isagoge*, presents the central headings of the classic Porphyrian tree as they appear in Figure 2. Such trees make it possible to construct definitions on the pattern of Aristotle: a species is defined according to its next highest type (the *genus proximum*), together with the specific traits which constitute the species (the *differentia specifica*). The stock example is still the definition of 'human being' as 'rational animal', citing both the proximate genus ('animal') and the specific difference that distinguishes human beings from animals of other kinds ('rational').

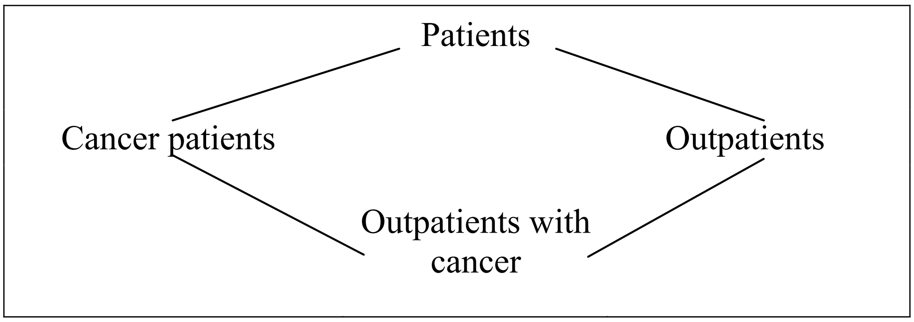
Figure 2: *The Structure of a Porphyrian Tree*



In information science, such tree structures are types of structured graphs. They flow in one direction, and the trees have a stem, the *genus ultimum*, from which increasingly finer branches split off, that finally end in the leaves or *species*. Taken together, all the ultimate kinds form the top-level ontology of an information system. In our stock example, the ultimate genus from which the species of human beings finally derives is normally assumed to be the category of substance or independent continuant (see chapter 8). Each element in such a tree (every node of the graph) has a unique supertype.

If we try to turn a cross-classification like the NCIT into a graph of this sort, then we face two problems. First, the uniqueness of a term's supertype is lost. Outpatients with cancer belong both to the supertype 'cancer patient' and to the supertype 'outpatient'. The branches of such a diagram no longer flow in a single direction. One element of the diagram can have multiple subtypes as well as multiple supertypes. Such situations are called *multiple inheritance* cases, since they allow us to produce diamond-formed structures like the example in Figure 3, in which the properties of the entities referred to by terms higher up in the hierarchy are inherited by the entities referred to by terms lower down along two or more distinct roots.

Figure 3: *An Example of Multiple Inheritance*

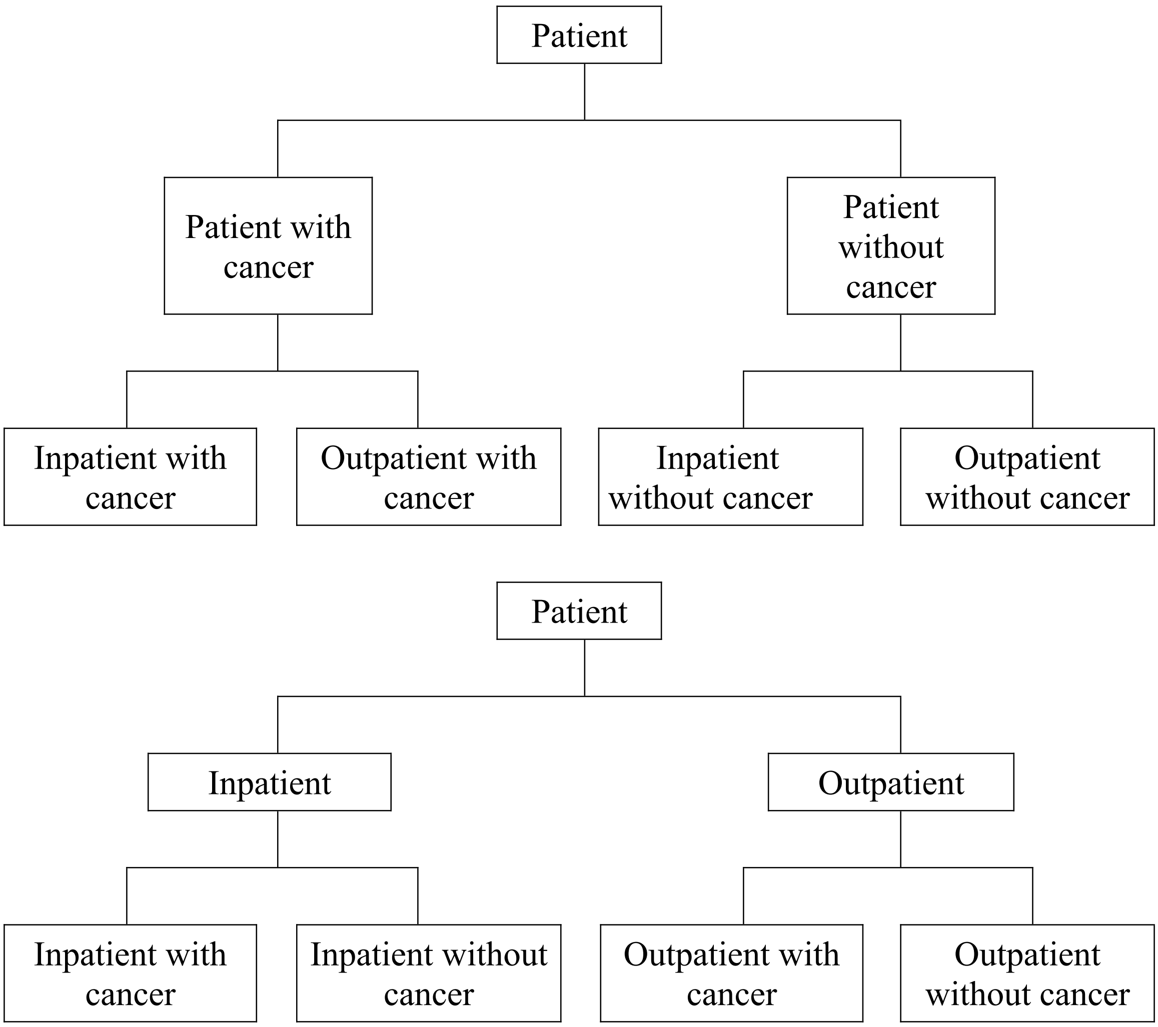


The second problem we face in such a situation is that, in order to construct a tree diagram after the fashion of Figure 4, we must determine which of these two traits should be considered prior in our classificatory hierarchy. In our classification, should we give priority the fact that the patient is an outpatient, or to the fact that he has cancer? To achieve a tree-structure, we must choose between the two.

Our choice between these two options would most likely be irrelevant to medical practice. But from the philosophical point of view, and from the point of view of ensuring consistency between different information systems (for example, in different medical specialties) such arbitrariness – and, thus, the possibility of making a random decision – is an unwelcome phenomenon, compounded by the fact that errors often result when distinct specification factors are combined within a single tree (Smith and Kumar, 2005). A cross-classification is based on the existence or nonexistence of two traits which are independent of one another. In the case of the patients in the NCIT, these are the questions: (1) for what is the patient being treated? (2) Is the patient staying overnight at the hospital?

The first question concerns the *reason* for the treatment, the second concerns an aspect of the *way in which he is treated*. Though both questions are important for the doctor in the hospital, each answer comes from totally different categories (as we will see in Chapter 8 of this book), and should be strictly distinguished in a classification system.

Figure 4: *Two Alternative Tree Diagrams*



One possibility for separating these aspects of a patient from one another is to create a multi-dimensional (or multi-axis) classification system. This approach is used, for example, by SNOMED CT, the *Systematized Nomenclature of Human and Veterinary Medicine*, developed by the College of American Pathologists (see SNOMED). In its third version, SNOMED distinguishes eleven different axes (or traits by which to classify), which can be combined with 17 qualifications. Figure 5 lists some of SNOMED's semantic axes.

Not every disease representation requires each of these axes. But by appealing to multiple axes, an encephalitis virus in a forest ranger can be coded as: TX2000 M40000 E30000 J63230 where the part of the code beginning with 'T' specifies the location of the disease, the part beginning with 'M' the body part affected, the part beginning with 'E' the cause of the disease (the virus), and the part beginning with 'J' the profession of the patient.

Figure 5: *Multi-dimensional Classification in SNOMED II*
(Dugas and Schmidt, 2003, 80)

Which morphological structure?	Morphology	M
Where is it situated?	Topography	T
What caused it?	Etiology	E
What is its effect?	Function	F
Which disease?	Disease	D
Which procedures have been applied?	Procedure	P
Connected with which profession?	Job	J

This correspondence of classificatory axes to kinds of questions is anticipated in the work of Aristotle, who uses terms for his categories which are taken mainly from interrogatory pronouns (Kahn, 1978, 227-278; cf. also next chapter).

2.3. *Uniformity: Laboratory Animals*

To classify patients according to both their cancer diagnosis and their status as outpatient leads to problems, not only with disjunctiveness and exhaustiveness, but it also violates the uniformity rule. Such a classification brings together distinguishing marks from different areas. This sort of violation is even more clearly manifest in the classification of laboratory animals in the NCIT. The importance of laboratory animals in cancer research is reflected in the variety of the twelve subtypes under the NCIT heading 'Laboratory_Animals'. Some of these types reflect particular things that have happened to the animals in question. For example, according to the NCIT definition, a 'Genetically_Engineered_Mouse' is a 'mouse that has been genetically modified by introducing new genetic characteristics to it'. Here, a DNA manipulation is given as the essence of a 'Genetically_Engineered_Mouse'. Other types, like 'Control_Animal', reflect a certain role the animals take on within a certain experimental design:

> **Control_Animal** NCI-GLOSS: the animals in a study that do not receive the treatment being tested. Comparing the health of control animals with the health of treated animals allows researchers to evaluate the effects of a treatment more accurately.

These definitions also draw on distinguishing marks that belong to quite different categories, namely natural kinds, roles, and being the subject of a procedure. Such categorial distinctions should be honored in a well-constructed ontology.

2.4. *Meta-Types and* '*Other*'

The NCIT is also deficient with regard to explicitness and precision. Like the CAT, the NCIT contains the entry 'Other'. This is a subtype of 'General_Modifier' (which is a subtype of 'Qualifier' that, in turn, is a subtype of 'Properties_and_Attributes') and is defined as 'Different than the one(s) previously specified or mentioned'. In all, there are approximately 80 *other*-involving entries in NCIT including for example: 'Carcinoma, Other, of the Mouse Pulmonary System'.

Another trait the NCIT shares with CAT is that of including meta-types (types that are dependent on the classification of which they are a part) alongside types within its hierarchy. For example, NCIT contains the type 'NCI-Thesaurus_Property', which is a subtype of 'Property' and is defined as a 'specific terminology property present in the NCI Thesaurus'. Meta-types even occur at the top-node level of the NCIT: its top-level features the heading 'Retired_Concept', defined as: a 'Concept [that] has been retired, and should not be used except to deal with old data'. This entry clearly mixes properties of the term with properties of the entities to which the term refers (compare Frege, 1884, § 53, and 1892, 192-205). Although it is undoubtedly useful to have a record of a term's properties, these properties should not be dealt with as if they were characteristics that a thing must have in order to instantiate a certain universal.

3. *Restrictive Conditions for Classifications*

In criticizing Borges's CAT and the NCIT, I have been guided by a vision of an ideal classification. According to this ideal, a classification consists of classes that are jointly exhaustive and pairwise disjoint (JEPD) and constructed out of ontologically well-founded distinguishing characteristics. There are a number of reasons why real-life classifications deviate from this ideal image.

A first group of limitations on classification derives from the domain to be classified. Particularly in the case of biological kinds, we have the problem that there is a large number of, for example, animal or plant or protein kinds which have not yet been scientifically described or even discovered. In addition, new genetic methods are enabling scientists to discover distinctions between kinds that are not available to traditional phenotype-based methods. The sheer number of kinds guarantees that biologists will have their work cut out for them for the foreseeable future. The number of animal kinds, alone, is estimated at approximately 30 million. There may be areas, such as human anatomy, that are close to being perfectly understood. But other areas are subject to constant growth in knowledge, such as zoology, botany, and especially genetics, which, because of the amount of available data, would hardly be possible to organize without the support of computers. Above all, however, we must bear in mind the likelihood of new species being discovered; not least because new species are constantly coming into existence. Such considerations, relating specifically to the domain to be classified, pose strict limitations on the exhaustiveness of a classification system. Some domains pose more principled problems for classification. Since, for example, bacterial genes can be switched from one bacterium to another and, because of the high rate of bacterial reproduction, can undergo rapid change, it is particularly difficult to distinguish stable species and kinds of bacteria.

A second group of limitations on classification derives from the technical side of the creation and application of classification systems. It does not matter whether we are dealing with a traditional, printed format, or a computer database; in either case, storage space is finite. Should computer programs be used for automated reasoning with the data contained within a classification, we have the problem of computability in addition to the problem of storage space. The time required for computation grows with the total number of classes, and with the number of inter-class relations with which a program must deal. Also, depending upon the programming language and its underlying logic and expressive power, there is the danger that a given task might not even be computable at all.

In addition to limitations posed by the domain of classification and by hardware and software, there are limitations posed by the human user. For while it is becoming ever easier and cheaper to extend the storage space on computers, the cognitive abilities of their human users have narrow limits. Human archivists and librarians are advised to use no more than one thousand systematically ordered key words (approximately) to index books or documents (Gaus, 2003, 93-94). Computers can, of course, use many more terms than this; the *NCI Thesaurus* with its 36,000 words is not a particularly large terminology database. As early as 2001, for example, the *Unified Medical Language System* (UMLS) encompassed 1.9 million expressions with more than 800,000 distinct meanings (see Dugas and Schmidt, 2003). But it is human curators who construct and maintain such artifacts, just as it is humans who later use them. The curators are experts who often specialize in the development of this particular kind of knowledge representation. But when, say, a general practitioner uses a certain classification as a diagnostic coding system in the process of billing, we have to ask how many diagnostic codes we can reasonably expect to be used in everyday practice.

Thus, there are several explanations for the deviation of real-life classifications from our envisioned classificatory ideal, and the main reason is that there are certain trade-offs between our various goals. If we want a complete representation of a given scientific domain, this might be far from a system that is easily comprehensible for a human user. If achieving completeness means to amass large amounts of data and to encode many relations between classes, we may also run into problems of computability. If, on the other end, we use simplifying types like *other* or *not otherwise specified*, we may run into trouble when updating the classification; for in the different versions *other* may have a quite different meaning and, thus, a different extension. But if we refrain from using other-types and simply give up the JEPD criterion, we lose a considerable amount of inferential strength. For, then, we no longer know that an entity that belongs to a supertype also belongs to one of the respective subtypes, and so on.

4. *Reference Ontologies*: *A Possible Solution*

A recent suggestion to solve this dilemma is based on a clear division of labor. We simply need two kinds of systems: reference ontologies and application ontologies. *Reference ontologies* should be developed without any regard to the problem of storage and the processing time, and they should represent, at any given time, the state of knowledge of the respective scientific discipline from which they derive (see OBO, 2006):

> A reference ontology is analogous to a scientific theory; it has a unified subject-matter, which consists of entities existing independently of the ontology, and it seeks to optimize descriptive or representational adequacy to this subject matter to the maximal degree that is compatible with the constraints of formal rigor and computational usefulness. Because a reference ontology is analogous to a scientific theory, it consists of representations of biological reality which are correct when viewed in light of our current understanding of reality (and thus it should be subjected to updating in light of scientific advance).

An *application ontology*, on the other hand, is analogous to a technical artifact like a computer program. Up to now, it was customary to build new ontologies from scratch for each new kind of application. This causes much trouble for anyone who wants to exchange or compare data among these different systems. It is better to use an already-existing reference ontology, from which we can derive the application ontology through a choice or combination of types from the reference ontology that are appropriate to the respective aim of the application ontology. Then, several such application ontologies can be mapped to each other through their respective reference to a common reference ontology.

While the task of maximally adequate representation of reality is transferred to the reference ontology, the application ontologies are constructed in light of the limitations posed by storage space, processing time, and the needs of the human users. While reference ontologies care about scientific virtues like completeness and precision, application ontologies care about engineering virtues such as efficiency and economic use of resources. The scientists of the OBO Foundry (see Smith, *et al.*, 2007) regard this as decisive progress:

> The methodology of developing application ontologies always against the background of a formally robust reference ontology framework, and of ensuring updating of application ontologies in light of updating of the reference ontology basis, can both counteract these tendencies toward ontology proliferation and ensure the interoperability of application ontologies constructed in its terms. (OBO, 2006)

5. *Exotic Thinking or Unfit Tool?*

Some philosophers have joined with Foucault in claiming that Borges's CAT possesses a certain exotic charm (Foucault, 1970; see also Jullien, 1990). I have shown that CAT is charming indeed, in that it can illustrate a wide range of possible mistakes in constructing taxonomies. CAT is, of course, literature and not science. As a contribution to science, it would not be evidence of exotic thinking, but rather of impractical thinking. For its part, the NCIT is not a piece of literature but is intended to be a piece of science. And it is, we believe, an example of very impractical thinking. In fact, the National Cancer Institute which maintains the NCIT is indeed itself dissatisfied with the present state of its thesaurus and its purported exotic charm, and is taking steps to improve it. As I have shown, such emendation is an excellent proof that technical applications can be helped by being built on foundations laid by philosophy.