Formalizing Botanical Taxonomies*

Uta Priss

School of Computing, Napier University, u.priss@napier.ac.uk

Abstract. Because botanical taxonomies are prototypical classifications it would seem that it should be easy to formalize them as concept lattices or type hierarchies. On closer inspection, however, one discovers that such a formalization is quite challenging to obtain. First, botanical taxonomies consist of several interrelated hierarchies, such as a specimen-based plant typology, a name hierarchy and a ranking system. Depending on the circumstances each of these can be primary or secondary for the formation of the taxonomy. Second, botanical taxonomies must comply with botanical nomenclature which follows a complicated rule system and is historically grown. Third, it may be difficult to design a formalization that is both mathematically appropriate and has a semantics which matches a taxonomist's intuition. The process of formalizing botanical taxonomies with formal concept analysis methods highlights such problems and can serve as a foundation for solutions.

1 Introduction

Biological taxonomies are often regarded as prototypical examples of classifications. First, it is usually assumed that taxonomic classes are defined using precise characteristics in the form of biological criteria that can be measured or in the form of specimen sets that are precisely grouped. Second, classes are usually non-overlapping with clear boundaries. Characteristics are usually non-gradual, i.e., they are either true or not true for an object that is to be classified. There is no grey-zone where a characteristic might apply only to a certain degree. Third, the classification usually forms a tree-hierarchy.

Given this assumed nature of biological taxonomies, it should be straightforward to formalize them as concept lattices in the sense of formal concept analysis. On closer inspection, however, one discovers that even biological taxonomies are more difficult to formalize than it might appear on first sight. This is demonstrated using the example of botanical taxonomies in this paper.

Botanical taxonomies are historically grown objects, which do not necessarily follow a simple construction principle. There are a variety of influences and contributing systems. For example, there are several separate hierarchies that contribute to the formation of a botanical taxonomy. These are a specimen-based plant typology (also called a "classification"), a name hierarchy and a ranking system. Each of the three is described

* This is a preprint of a paper published in De Moor; Lex; Ganter (eds.), Conceptual Structures for Knowledge Creation and Communication. Proceedings of the 11th International Conference on Conceptual Structures, Springer Verlag, LNAI 2746, 2003, p. 309-322.

in section 2. The notion of “taxonomy” usually refers to a combination of a classification with a name hierarchy. A taxonomy consists of “taxa” (singular: “taxon”).

Section 3 provides a short introduction to the basic terms of formal concept analysis. But it is assumed in this paper that the reader is already somewhat familiar with formal concept analysis. If not, the reader is referred to Ganter & Wille (1999). Section 4 provides a short description of the Prometheus Project (Pullan et al., 2000), which is based on an object-relational modeling of taxonomies. Section 5 discusses some of the advantages and challenges involved in formalizing taxonomies with formal concept analysis. Section 6 considers the dynamic nature of botanical taxonomies and the notion of “closure systems”. Section 7 provides the details of the formalization of taxonomies with formal concept analysis and Section 8 discusses some problems involved in applying a ranking system to a combination of multiple taxonomies.

2 Three Hierarchies in Botanical Taxonomies

A specimen-based plant typology (or classification) is created by taxonomists who group the specimens that they are interested in according to characteristics that can be identified for these specimens. This corresponds to a formal context in formal concept analysis which has specimens as objects and characteristics as attributes. (Although the ranking system must be included in such a formal context and it is actually not practical to use characteristics as attributes. See below for details.) One problem is that because of the size of modern taxonomies and the fact that usually many taxonomists contribute to a classification, it is difficult to oversee a single classification. Other problems are mentioned by Pullan et al. (2000) who state that classifications must cope with “historical data, newly described taxa, new revisions and conflicting opinions”.

The second component of taxonomies is a name hierarchy. Name hierarchies are based on a strict nomenclature, “the International Code of Botanical Nomenclature” (Greuter et al., 1994)), which is called “the Code” in the rest of this paper. Unfortunately, because both name hierarchies and the nomenclature are historically grown and the Code is a very complicated rule system, there is no simple mapping from any specimen-based plant typology to any name hierarchy. Names are only valid within the context in which they are published. That means that a single name can refer to different specimens in different taxonomies. For example, the family of “lilies (Liliaceae)” includes “asparagus” in some taxonomies, whereas in others it does not. Furthermore, a single specimen may have different names in different taxonomies. Names above “genus” are formed differently from names below or at genus level. Names depend on rank. For example, the same specimen “Eriogonum umbellatum var. subaridum” at the variety rank, is called “Eriogonum umbellatum subsp. ferrisii” at the subspecies rank and “Eriogonum biumbellatum” at the species rank (Reveal, 1997). There are also rules of historical authorship. Usually the first published name for a taxon should be used. But there are cases where publications are “forgotten” and then rediscovered (Reveal, 1997). In this case, the Code may specify exceptions to the normal rules and state explicitly which names are to be used in preference.

The ranking system contains the ranks “variety”, “species”, “genus”, “family” and so on. The same ranking system is used in principle for all taxonomies. That means

that no taxonomist can invent new ranks. But not all taxonomies must have all ranks. For example, the rank of "subspecies" may not always be used. As mentioned above, names depend on ranks. But even classifications depend on ranks. The lowest common supertype of two specimens can occur at different ranks in different classifications. Furthermore, the same specimen set can be used at different ranks. For example, if a family has only one genus, then the specimen set of that family and of that genus are identical. But the taxon for this family and this genus are different and will have different names assigned. Thus a classification cannot be generated from just the specimen sets. The rank information is essential. While it is relatively easy to define ranks in a tree hierarchy, assigning ranks to a multi-hierarchy or lattice is more difficult because there could be paths of different length from the top element to any taxon. Thus if different taxonomies are combined, the ranking system becomes problematic. But because they are used both in the classification and in the naming hierarchy, ranks cannot be ignored.

3 Formal Concept Analysis

This section provides a short introduction to the basic notions of formal concept analysis which are relevant for taxonomies. Formal concept analysis (Ganter & Wille, 1999) provides a formal framework and vocabulary suited for classification systems, taxonomies, ontologies and hierarchies. Its main unit of study are concepts within a conceptual hierarchy. In the case of type hierarchies, concepts are called "types"; in classification systems they are called "classes"; in ontologies they are called either "concepts" or "types" or "classes"; and in taxonomies they are called "taxa". Concepts are to be distinguished from objects and attributes, which are usually formal representations of elements of some domain. Objects can also be called elements, individuals, tokens, instances or specimens. They describe the extensional aspects of concepts. Attributes can also be called features, characteristics, characters or defining elements. They describe the intensional aspects of concepts. Both extensional and intensional elements can be used separately or together to describe or define concepts. Alternatively, concepts can be defined by establishing relations with other concepts. Thus there are multiple ways of describing, characterizing or comparing concepts. Formal concept analysis provides a formal framework in which all of these can be combined and applied.

There are two primary views of the data in formal concept analysis: formal contexts and concept lattices. Formal contexts are represented as binary relational tables. A formal context (O, A, I) consists of a set O of objects, a set A of attributes and a binary relation I between them, i.e., $I \subseteq O \times A$. Instances of the relation are either denoted by oIa or by $(o, a) \in I$ for an object $o \in O$ and an attribute $a \in A$. They are both read as "object o has the attribute a". Two mappings are defined: $I(o) := \{a \in A | (o, a) \in I\}$ provides for each object the set of its attributes and $I(a) := \{o \in O | (o, a) \in I\}$ provides for each attribute the set of objects that have that attribute. If one applies this mapping repeatedly, one will find that after two iterations there are no further changes. That is if one starts with a set of objects, takes all their attributes and then looks up all other objects that also have these attributes one can stop at exactly that point. Further applications of $I()$ will not add further objects or attributes. Formally, $\forall_{o \in O, a \in A} : I(I(I(o))) = I(o)$ and $I(I(I(a))) = I(a)$.

Concepts are pairs of sets of objects and sets of attributes of a formal context that do not change if $I()$ is applied. Formally a pair (O_1, A_1) with $O_1 \subseteq O$ and $A_1 \subseteq A$ is a concept if $I(O_1) = A_1$ and $O_1 = I(A_1)$. The set of objects of a concept c is called its *extension* and denoted by $ext(c)$. The set of attributes of a concept is called its *intension* and denoted by $int(c)$. An ordering relation is defined on a set of concepts as follows: a concept c_1 is a subconcept of another concept c_2 if $ext(c_1) \subseteq ext(c_2)$. This is equivalent to $int(c_2) \subseteq int(c_1)$. For example, "Bellis (bellflower)" is a subconcept of "Asteraceae" if all instances of Bellis are also instances of Asteraceae or, equivalently, all attributes of Asteraceae also apply to Bellis. This conceptual ordering is denoted by $\leq$. The set of concepts with its ordering, $(C, \leq)$, is not only a partially ordered set but even a (concept) lattice, which means that for each set of concepts there exists a unique infimum (largest common subconcept) and a unique supremum (largest common superconcept). Infimum and supremum of concepts are denoted by $c_1 \wedge c_2$ and $c_1 \vee c_2$, respectively.

There are many applications for formal concept analysis. To convert a traditional tree-hierarchy into a concept lattice, one must formally add a bottom concept which has an empty extension and has all the attributes of the underlying formal context in its intension. In some applications only either sets of objects or sets of attributes are of interest. It is sufficient to either only specify the extensions of a concept lattice or only the intensions to generate the lattice. The intensions (or extensions, respectively) can then be formally labeled, such as A_1, A_2, and so on, but have no further meaning for the domain. It should be pointed out that concepts, suprema, infima, the conceptual ordering and so on all depend on the formal context in which they are defined.

4 The Prometheus Project

The Prometheus project (Pullan et al., 2000) provides an object-oriented formalism for taxonomies that is implemented in an object database system (Prometheus, 2002). This implementation improves on prior attempts at implementing taxonomies which according to Berendsohn (1997) often suffer from "a frequent error of over-simplification of taxonomic data by non-taxonomist database designers" (quoted after Pullan et al. (2000)). A first main achievement of Prometheus is a clear separation of the specimen-based classification and the nomenclature-based name hierarchy. Apparently prior implementations of taxonomic databases failed to separate these two thus making it impossible to combine more than one taxonomy in a single database. Prometheus recognizes the value of combining several different taxonomies and purposefully designs the separation of classification and name hierarchy to achieve this combination. It also provides an impressive graphical interface that facilitates simultaneous browsing through multiple taxonomies (Graham, 2001).

A second main achievement of Prometheus is a concentration on the specimens as the main drivers for the classification. Biological characteristics are not as unambiguous as one might naively assume (Pullan et al, 2000). In fact characteristics, such as "hairy leaves", are very imprecise. Furthermore, they are not usually checked for all specimens in the database but only for those within a certain smaller subset. Some taxonomists might even assign characteristics based on a prototypical image that they have

of the specimens in a set, which may not be true for all actual specimens. Thus biological characteristics are not attributes in the sense of formal concept analysis. The Prometheus project defines the "circumscription" not in the traditional manner as referring to the characteristics of a taxon but instead as the set of specimens of a taxon. Thus in formal concept analysis terms, the taxonomy is entirely extensionally defined. This ensures that both taxa and circumscriptions are precisely defined and that different taxonomies can be combined and compared based on the specimen sets. The potential problem of having to incorporate too much data (if all specimens of the world were included in a taxonomy) is overcome by focusing on "type specimens", i.e., such specimens that actually contribute to the distinctions made in the classification. In formal concept analysis terms this corresponds to using a "purified" formal context.

5 Advantages and Challenges Involved in Formalizing Taxonomies with Formal Concept Analysis

There are some shortcomings of the object-oriented modeling as used in the Prometheus Project which a formal concept analysis modeling can clarify. It is not discussed in this paper whether or not a formal concept analysis-based implementation would improve the performance of the database. That may or may not be the case. But it is claimed that a formal concept analysis-based representation simplifies the terminology and the conceptual understanding of the problems involved in modeling taxonomies. There is, however, at least one definite practical advantage of using formal concept analysis modeling instead of object-oriented modeling, which is that the Prometheus software had to be specifically designed for its purpose. This required a significant amount of resources. Most likely Prometheus cannot easily be used for other types of hierarchical data without major redesign. On the other hand, formal concept analysis software provides general multi-purpose tools for dealing with hierarchical data. Each new project requires some specific adjustments but the required time and effort is much smaller than designing a single-purpose object-oriented implementation.

The remainder of this section discusses some of the differences between object-oriented and formal concept analysis modeling. It also discusses the challenges that formal concept analysis models must overcome to be acceptable to application domain experts. For a start, domain experts often do not see a need for mathematical formalization at all. In the case of taxonomists, they already have a complicated semi-formal system in their Code. One advantage of the Prometheus model is that on the surface it is simpler. It matches fairly closely to the taxonomic terminology without introducing formal constructs that do not have an obvious reason for existence from a taxonomist's viewpoint. An example for such formal constructs is the bottom concept with an empty set of specimens that a formal concept analysis modeling adds to every tree-hierarchy.

Another formal construct is what Berendsohn (1997) calls "potential taxa" which can be automatically generated higher up in the hierarchy, Pullan et al. (2002) express discomfort with these, which they call "informationless" taxa. They state that Prometheus should not extrapolate "far beyond what can be supported by the original sources". From a formal concept analysis viewpoint automatically generated information does not pose a problem as long as users are aware of what is input to the system

and what is generated. But supporting such awareness and ensuring that users feel comfortable with the system can be difficult. Thus a formal concept analysis model is more complicated on the surface than an object-oriented formalism because it contains these structures which Pullan et al. call “informationless”. But it is claimed in this paper that the “deeper level” structures, i.e., underlying mathematical formalisms and formal semantics, of a formal concept analysis modeling are actually simpler.

There is no detailed mathematical formalism with formal semantics provided for Prometheus. But it can be foreseen that such a formalism would be complicated. An advantage of a simpler underlying mathematical formalism is that it is easier to be checked for consistency, completeness and semantics. An extreme example of a simple surface model which completely lacks an underlying formalism is the Code itself. Due to its lack of mathematical formalism it is ambiguous as it is. It would be very difficult to model it precisely and to check it for consistency.

It has already been mentioned that formal concept analysis allows to reuse its software for different hierarchy-related projects instead of having to write ad-hoc solutions for each project. A similar argument can be made on the level of mathematical structures. Formal concept analysis builds on complex but well known mathematical structures, such as lattices. If it can be shown that taxonomies can be represented as lattices in the sense of formal concept analysis, then it follows that numerous statements about formal properties can automatically be inferred about taxonomies. Formalizations only have a purpose if they support such inferences. There is a chance that inferentially derived knowledge about taxonomies contradicts domain knowledge. While such discrepancies could be caused by inappropriate use of formalizations, they can also point to problems within the domain knowledge. Philosophers, such as Peirce or Brandom (1994) emphasize the use of logic (formalisms and inferences) for the purpose of “making it explicit”. This paper claims that formal concept analysis can serve as an inferential tool that explicates structures in taxonomies. For example, it will be shown below that the notion of “rank” causes problems at least in the case of a combination of multiple taxonomies. Maybe taxonomists should review their use of “rank”.

Another advantage of formal concept analysis is the explicit use of “context”. A consequence of a lack of considering context, is a confusion between member-of and subset-of relations. For example, CYC (1997) explains the differences between what they call the isa and the genls relation and the resulting problems with transitivity. From a formal concept analysis viewpoint, a concept is a subconcept of another concept if the extension of the first concept is contained in the extension of the second concept. The subconcept relation among concepts thus corresponds to a subset relation on the set of extensions. The subconcept relation is transitive. The member-of or element-of relation holds between objects and concepts. If an object is an element of a concept than it is also an element of all its superconcepts. But an object is always at the left-hand side of the member-of relation. Nothing can be an element of or a subset of an object. At least not within the same context. Concepts are never elements of other concepts within the same context. But a concept in one context can correspond to an object in another context which is then an element of other concepts in that second context.

For example, “Asteraceae” is the name of the family of asters. In a context which has plant specimens as objects and contains a concept “Asteraceae”, certain specimens

can be elements of “Asteraceae”. In a different context, “Asteraceae” can denote an object, which can then be an element of a concept called “family”. Thus the statements “this specimen is an element of Asteraceae” and “Asteraceae is a (or is an element of) family” can both be true, but not in the same context! The false statement “this specimen is a family” is not implied because inferences across contexts are not allowed without further constraints.

For Prometheus the distinction between element-of and subset-of is important because a form of “element-of” is used in classifications to describe the assignment of specimens to taxa. Two forms of “subset-of” are used in name hierarchies: first, there is the “placement” of a taxon within a higher-ranked taxon. Second, there is the “type definition” which for each taxon points to a taxon of lower rank or to a specimen set. The reason for the need of type definitions is that the name of a higher level taxon is formed based on one prototypical specimen or lower level taxon. For example, the family “Asteraceae” is named after the prototypical genus “Aster”. Going up in the hierarchy, “Asteraceae” provides the basis for “Asterales”, which in turn provides the name for “Asteridae”. But the notion of “type definition” can lead to a confusion of “element-of” and “subtype-of” because either specimens or subtaxa are used. A solution would be to formalize name hierarchies without referring to specimens at all, but instead use names as basic elements, ignoring whether they are names of specimens or names of taxa.

6 Dynamic Taxonomies and Taxonomy Merging

Several further aspects of taxonomies should be mentioned before the formalization is explained in detail. Taxonomists have an implied dynamical view of classification. Because new specimens are discovered, existing classifications are revised, and new scientific knowledge is gained through new methods (such as genetic analysis), taxonomic knowledge is continuously changing. A certain static view of taxonomies is imposed by distinguishing between published taxonomies and taxonomies that are “in-progress”. Published taxonomies never change because if revisions of a taxonomy are published, then this is viewed as a new separate taxonomy. On the other hand, working taxonomists have their own personal taxonomies “in-progress” which they can be changing on a daily basis. But these are not shared beyond an immediate circle of colleagues. As long as a taxonomy is changing it is private and not published. Name hierarchies are attached to published taxonomies. Taxonomies that are in-progress normally consist of classifications, not name hierarchies. In Prometheus, the names can be automatically calculated to avoid naming errors on the taxonomist’s part.

A second aspect to be mentioned is the notion of a “closure system”, which may be the minimal formal structure that should be considered for combining taxonomies and classifications. Viewing a specimen-based classification as a closure system means that if sets of specimens are intersected these intersections are also part of the classification. For example, a gardener might use both a taxonomy to obtain information on plant families and a plant catalog, which provides information on plant habitat and hardiness. A plant catalog is not a taxonomy in the sense of the Code, but it can be considered what happens if it is combined with a taxonomy. The plant catalog may provide sets

of specimens, such as “spring flowering plants”, which can be intersected with specimen sets from a taxonomy to form sets such as “asters that flower in spring”. These are meaningful sets which are easily obtained from the combination. Thus it is reasonable to demand that intersections are calculated when classifications or taxonomies are combined. Of course, only those intersections are added to the merged classification that provide new information. Since roses flower in summer there will not be a new class “summer-flowering roses”. Instead the specimen set for “roses” will be contained in the specimen set for “summer-flowering plants”. Thus if classifications are merged that are very similar, not many new interesting intersections may occur.

Unions of specimen sets, however, are usually not considered at all. For example, a set of plants that are either asters or spring-flowering is an odd combination. While unions of specimen sets need not be included in a classification, it is important to know which is the smallest set that contains a union. This is called a *closure set*. For example, while a union of asters and roses may not be a meaningful set, it is important to know which is the smallest set that does contain both asters and roses because that is the point where the upwards paths from aster and rose meet in a classification. In the case of asters and spring flowering plants the closure set is probably the set of all plants, which is not very interesting but by default contained in a classification. A *closure system* thus provides intersections and closure sets for all sets of specimens and is an intuitive formalization of classifications. Tree hierarchies are automatically closure systems if the empty set is formally added as the shared bottom node of the tree. Formal concept lattices are closure systems with additional structure.

7 The Formalization of (Botanical) Taxonomies

7.1 Specimens, Taxa, Ranks

In what follows S denotes a set of specimens; T denotes a set of taxa; and R denotes a set of ranks. It should be noted that all sets are finite and $S \cap R = \emptyset$, $S \cap T = \emptyset$, $T \cap R = \emptyset$.

7.2 Closure Systems

Let $P(S)$ denote the power set (or set of subsets) of S and $\overline{S}$ a subset of $P(S)$. According to standard order theory, a set $\overline{S}$ forms a *closure system* if

$$S \in \overline{S} \text{ and } (\overline{S}_1 \subseteq \overline{S} \Longrightarrow \bigcap \overline{S}_1 \in \overline{S}) \tag{1}$$

Also according to standard order theory it follows that if $\overline{S}$ is a closure system then $\overline{S}$ with $\subseteq$ forms a complete (concept) lattice

$$(\overline{S}, \subseteq) \text{ with formal context } (S, \overline{S}, \in) \text{ and } \top := S \text{ and } \bot := \bigcap \overline{S} \tag{2}$$

7.3 Tree Hierarchy

Tree hierarchies are closure systems if a bottom node is formally added. They are defined as follows:
A set $\overline{S}$ with $\overline{S} \subseteq P(S)$ forms a *tree hierarchy with bottom element* if

$$S \in \overline{S} \text{ and } (\forall_{a,b \in \overline{S}} : a \not\subseteq b \text{ and } b \not\subseteq a \Rightarrow a \cap b = \emptyset) \tag{3}$$

If a tree has at least two incomparable nodes and $\emptyset \in \overline{S}$ then $\overline{S}$ is a closure system and $(\overline{S}, \subseteq)$ is a complete lattice with top element $\top := S$ and bottom element $\bot := \emptyset$.

It is straightforward to interpret a specimen-based plant typology as a concept lattice in the sense of formal concept analysis. But this is not yet sufficient for the treatment of specimen classifications because as discussed before these depend on rank. Two different taxa can have the same specimen set but different ranks. This is always the case, if a taxon at one rank has only one immediate subtaxon at a lower rank.

7.4 Rank Order and Rank

The ranking systems considered by taxonomists are always linear orders, thus only linear orders are considered here. Using specimens as objects and ranks as attributes would not yield a useful concept lattice because the property of "being at a certain rank" is not inherited. For example, members of "species" are not also members of "genus". But the property of "being at most at a certain rank" is inherited. Thus ranking systems can be formalized by using "at-least-ranks" as objects and "at-most-ranks" as attributes.

A partially ordered set $(R, \preceq)$ is called a *rank order* if $\preceq$ is a linear order. The elements of R are called *ranks*. With $R^{\geq} := R$ and $R^{\leq} := R$, a rank order can be modeled as a concept lattice

$$(R, \preceq) \text{ with formal context } (R^{\geq}, R^{\leq}, \vdash) \text{ and } r_1 \vdash r_2 \Longleftrightarrow r_1 \preceq r_2 \tag{4}$$

For example, "at-least-species below at-most-genus" if and only if "species below genus".

7.5 (Botanical) Taxonomies

The following two assumptions are necessary for the definition of a (botanical) taxonomy:

1) The set of sets of specimens $(\overline{S}, \subseteq)$ forms a closure system.

2) A map $rank : \overline{S} \rightarrow R$ assigns to each set of specimens in $\overline{S}$ exactly one rank with the condition

$$\forall_{S_1,S_2} \in \overline{S} : S_1 \subset S_2 \Rightarrow rank(S_1) \prec rank(S_2) \tag{5}$$

The assigned rank is the maximum rank at which a taxonomist wants to see this specimen set uniquely describing a taxon. A specimen set can describe different adjacent ranks if the taxonomist sees no need to divide the set into subsets at the lower rank. To formally distinguish these sets, each set of specimens is unioned with its rank,

i.e., $S_1 \cup rank(S_1)$. If $S_1 \subset S_2$, $rank(S_1) \prec r \prec rank(S_2)$ and there exists no S_3 with $rank(S_3) = r$, then both $S_2 \cup rank(S_2)$ and $S_2 \cup r$ are formed. The resulting set of specimen sets unioned with ranks is denoted by $\overline{S}^R$. The map $rank$ is extended to $\overline{S}^R$ by mapping each set to its rank. It follows that if the top rank is assigned to any specimen set, then $rank$ on $\overline{S}^R$ is surjective.

A *(botanical) taxonomy* is now defined as a concept lattice $(T, \leq)$ with the following formal context:
$(S \cup R, \overline{S}^R, I_{SR})$ with $I_{SR} = \{(s_1, S_1) | s_1 \in S_1\} \cup \{(r_1, S_1) | r_1 \vdash rank(S_1)\}$.

Because of $S \cap R = \emptyset$, this is a disjoint union and can be represented as $I_{SR} = I_{SR}|_S \cup I_{SR}|_R$. The elements of $(T, \leq)$ are denoted by $t \in T$ and are called *taxa*. The map $rank$ can be extended to all taxa as follows: $rank : T \rightarrow R$ with $rank(t) = max\{r \in R^{\geq} | r \in ext(t)\} = max_r ext(t)|_R$.

A taxon is uniquely characterized by its rank and the set of specimens it refers to. Figure 1 shows two examples of taxonomies. The terms in boxes represent names from a name hierarchy which are assigned to the taxonomies. From a formal concept analysis viewpoint, there is no difference between the dashed and the solid lines. From a taxonomist's viewpoint, the dashed lines are pointers from the ranks to the taxa at those ranks. In figure 1, "carrot" in the left taxonomy and "apple" in the right taxonomy represent examples of specimen sets that occur at two different ranks. On $\overline{S}$, these are mapped to rank 2; in $\overline{S}^R$ these correspond to {carrot, 1}, {carrot, 2}, {apple, 1}, {apple, 2}.

7.6 A Specimen/Rank Information Channel

A taxonomy as defined in the previous section can also be interpreted as an information channel (in the sense of Barwise & Seligman (1997)) between a specimen-based closure system $(S, \overline{S}, \in)$ and a rank order $(R^{\geq}, R^{\leq}, \vdash)$ (cf. figure 2). This corresponds to an intuitive notion that the specimen classification and the rank order are independent of each other but both contribute to the taxonomy.

The infomorphisms are constructed as follows:
a) With $sp : \overline{S}^R \rightarrow \overline{S}$ and $tax : S \rightarrow S$ as identity map it follows that
$s_1 \in sp(S_1) \Longleftrightarrow tax(s_1)\ I_{SR}\ S_1$.
b) With $rk : R^{\geq} \rightarrow R$ as an identity map and $rank : \overline{S}^R \rightarrow R^{\leq}$ as defined above it follows that
$r_1 \vdash rank(S_1) \Longleftrightarrow rk(r_1)\ I_{SR}\ S_1$ because of the definition of I_{SR}.

The map sp is extended to all taxa in the following manner: $sp : T \rightarrow \overline{S}$ with $sp(t) := \{s \in S | s \in ext(t)\} = ext(t)|_S$. This is well defined because $sp(t)$ is an element of $\overline{S}$ due to the fact that $I_{SR}|_S$ is equivalent to "$\in$" in $(S, \overline{S}, \in)$.

7.7 Properties of Taxonomies

The following 7 properties of taxonomies can be proved. These properties correspond to intuitive notions that taxonomists have about their taxonomies and demonstrate that the semantics of the formalization as derived so far does not contradict traditional taxonomic knowledge.

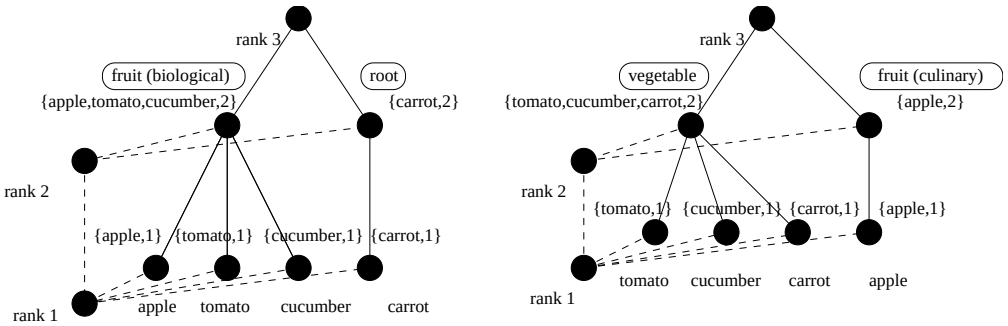

Fig. 1. Two taxonomies

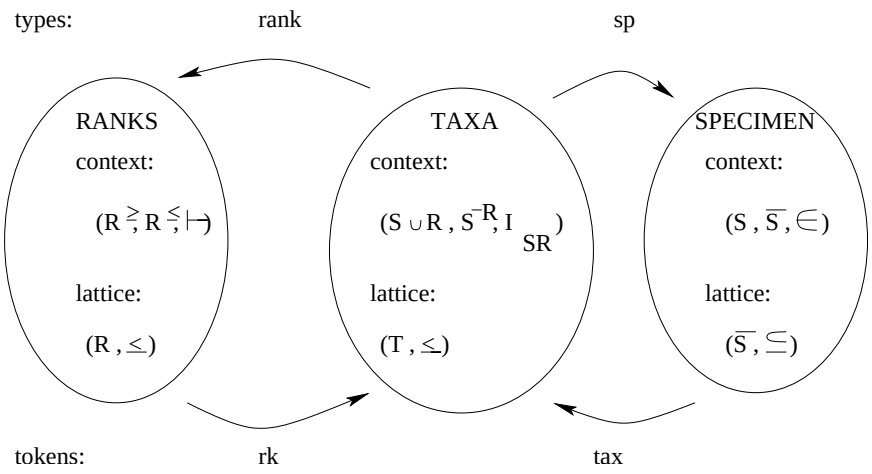

Fig. 2. A specimen/rank information channel

– 1) For each specimen exists at least one taxon so that $s \in ext(t)$.
– 2) For each specimen exists exactly one minimum taxon $tax(s)$ in $(T, \leq)$ so that $s \in ext(t)$, i.e., $tax(s) := \bigwedge\{t \in T | s \in ext(t)\}$.
  Proof: $(T, \leq)$ is a lattice.
– 3) For each rank exists at least one taxon so that $r \in ext(t)$.
  Proof: Due to the set of objects: $S \cup R^{\geq}$.
– 4) $t_1 \leq t_2 \iff sp(t_1) \subseteq sp(t_2)$ and $rank(t_1) \preceq rank(t_2)$.
  Proof: $sp(t_1) \subseteq sp(t_2)$ and $rank(t_1) \preceq rank(t_2) \iff$
  $ext(t_1)|_S \subseteq ext(t_2)|_S$ and $max_r ext(t_1)|_R \preceq max_r ext(t_2)|_R \iff$
  $ext(t_1)|_S \subseteq ext(t_2)|_S$ and $ext(t_1)|_R \subseteq ext(t_2)|_R \iff$
  $ext(t_1) \subseteq ext(t_2) \iff t_1 \leq t_2$
– 5) $t_1 = t_2 \iff sp(t_1) = sp(t_2)$ and $rank(t_1) = rank(t_2)$
  Because $\leq$ is ordering and $t_1 \leq t_2$ and $t_2 \leq t_1 \iff$
  $sp(t_1) \subseteq sp(t_2)$ and $rank(t_1) \preceq rank(t_2)$ and
  $sp(t_2) \subseteq sp(t_1)$ and $rank(t_2) \preceq rank(t_1)$
– 6) $t_1 < t_2 \iff sp(t_1) \subset sp(t_2)$ and $rank(t_1) \preceq rank(t_2)$ or $sp(t_1) \subseteq sp(t_2)$ and $rank(t_1) \prec rank(t_2)$
  Because: $t_1 \neq t_2 \iff sp(t_1) \neq sp(t_2)$ or $rank(t_1) \neq rank(t_2)$
  $t_1 \leq t_2$ and $t_1 \neq t_2 \iff sp(t_1) \subseteq sp(t_2)$ and $rank(t_1) \preceq rank(t_2)$ and
  $(sp(t_1) \neq sp(t_2)$ or $rank(t_1) \neq rank(t_2))$
– 7) $t_1 \wedge t_2 = t \iff sp(t_1) \wedge sp(t_2) = sp(t)$ and $rank(t_1) \wedge rank(t_2) = rank(t)$.
  $t_1 \vee t_2 = t \iff sp(t_1) \vee sp(t_2) = sp(t)$ and $rank(t_1) \vee rank(t_2) = rank(t)$.

8 Rank and Multiple Classifications

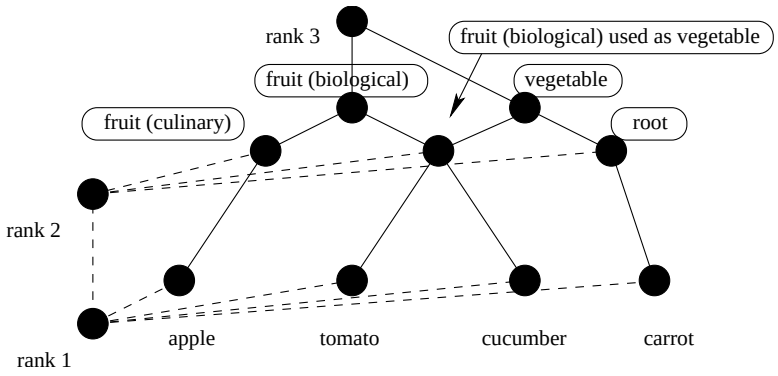

Fig. 3. Merging taxonomies

Botanical taxonomies are usually modeled as tree-hierarchies. But if multiple hierarchies are combined, then the resulting structure is a multi-classification, which is

not usually a tree-hierarchy. Intersections between specimen sets from different classifications emerge as new classes. The example in figures 1 and 3 shows a biological classification in which “tomato”, “cucumber” and “apple” are in a class called “fruit (biological)”. In a culinary classification, “tomato” and “cucumber” are grouped under “vegetable”, and “apple” under “fruit (culinary)”. In the merged classification new classes arise, such as “fruit (biological) used as vegetable” which contains “tomato” and “cucumber”. This new class emerges because it adds new information. Because all elements under “fruit (culinary)” are also under “fruit (biological)” there is no new class “fruit (biological) used as fruit (culinary)”.

A problem with multi-classifications is the meaning of “rank”. In a tree-hierarchy it is fairly easy to establish ranks because there is only one path from each specimen set to the top of the tree. In a multi-hierarchy there are possibly several paths of different length from a single taxon to the top of the classification. It is thus not possible to start at the bottom and assign ranks by following all paths that lead upwards. An obvious idea is to somehow combine the original ranks of the original tree-like classifications into a shared rank of the multi-classification. But it is a question which ranks to assign to the emerging classes. In the example, “fruit (culinary)”, “fruit (biological)” and “vegetable” are all of rank 2. What rank should then be assigned to “fruit (biological) used as vegetable”? It is not possible to choose an already existing rank that is smaller than 2 because “tomato” and “cucumber” are at rank 1. Another possibility would be to invent new ranks for the new emerging classes (such as rank 1.5) but botanical taxonomists are not allowed to invent new ranks.

Thus the only remaining possibility is for a class that emerges as an intersection of classes x_1, x_2, and so on to assign the maximum rank that is shared by all these classes. In the example above, that means that “fruit (biological) used as vegetable” receives rank 2 even though that is the same rank as its parent classes. In this approach, the main function of using ranks is to ensure that every class can be uniquely identified by its specimens and its rank. This approach is consistent with the definitions and propositions of section 7. But as the example in figure 4 shows it is not in general correct to assume that if a set of specimens of one class is contained in the set of specimens of another class that then the second class should be of higher or equal rank than the first class. That means that $sp(t_1) \subset sp(t_2) \not\Rightarrow rank(t_1) \preceq rank(t_2)$ and thus while equation (5) is true for the original specimen sets of a classification it cannot be generalized to all taxa.

9 Conclusion

This paper provides the foundation for a mathematical formalization of botanical taxonomies. The formalization shows that the notion of “rank” is problematic with respect to merged taxonomies and non-tree-like taxonomies. In these cases, ranks only serve to distinguish taxa but not to rank-order them. If “rank” cannot be generalized to these cases, then maybe taxonomists should review their use of ranks. Maybe ranks should be applied to taxonomies as a meta-level property. But then they could not be used to define taxa.

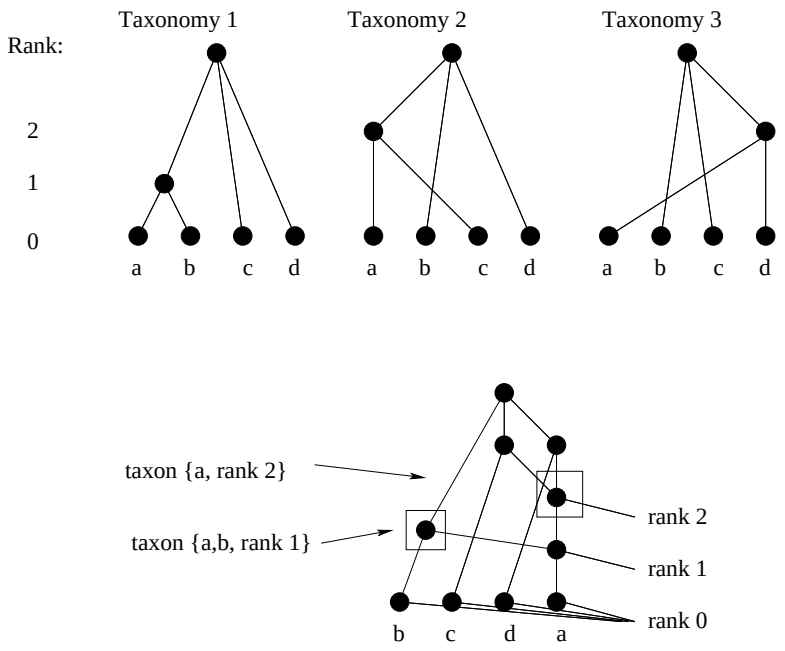

Fig. 4. The question of rank

For future research, it would be interesting to compare the process of botanical classification to similar historically grown classification tasks. Library cataloging employs a complex rule based system similar to the Code used in botanical taxonomies. A formalization of library cataloging rules might provide insights into botanical classification and vice versa.

Acknowledgments

I wish to thank my colleagues Jessie Kennedy, Andrew Cumming, Gordon Russell and Cedric Raguenaud for stimulating discussions on this topic.

References

1. Barwise, J.; Seligman, J. (1997). *Information Flow. The Logic of Distributed Systems.* Cambridge University Press.
2. Brandom, R. (1994). *Making It Explicit: Reasoning, Representing, and Discursive Commitment.* Cambridge, MA: Harvard University Press, 1994.
3. Berendsohn, W. G. (1997). *A taxonomic information model for botanic databases: the IOPI Model.* Taxon, 44, p 207-212.
4. CYC (1997). *A note about isa and genls.* Online available at http://www.cyc.com/cyc-2-1/footnotes.html#IsaGenls

5. Ganter, B.; Wille, R. (1999a). *Formal Concept Analysis.* Mathematical Foundations. Berlin-Heidelberg-New York: Springer, Berlin-Heidelberg.
6. Graham, M.; Kennedy J. (2001). *Combining linking & focusing techniques for a multiple hierarchy visualisation.* 5th International Conference on Information Visualisation - IV2001, University of London, London, IEEE Computer Society Press, p 425-432.
7. Greuter, W.; Barrie, F. R.; Burdet, H. M.; Chaloner, W. G.; Demoulin, V.; Hawksworth, D. L.; Jorgensen, P. M.; Nicholson, D. H.; Silva, P. C.; Trehane, P.; McNeill, J. (1994). *International Code of Botanical Nomenclature (Tokyo Code).* Regnum Veg. 131.
8. Pullan, M. R.; Watson, M. F.; Kennedy, J. B.; Raguenaud, C.; Hyam, R. (2000). *The Prometheus Taxonomic Model: a practical approach to representing multiple classifications.* Taxon, 49, p 55-75.
9. Prometheus (2002). On-line available at http://www.dcs.napier.ac.uk/ prometheus/
10. Reveal, J. (1997) *One and Only One Correct Name?*. On-line available at http://www.inform.umd.edu/PBIO/pb250/onename.html