4

Ontology creation

Summary. In the last chapter we discussed languages for explicating information semantics and argued for the need of an integration at the language level. We now draw attention to the nature and the content of ontologies needed to support information sharing. The goal is to define an architecture combining the advantages of global and local ontologies and to show how this infrastructure can be derived from an information-sharing task.

The acquisition of semantic knowledge has been identified to be a major bottleneck not only in information sharing but also in many other areas going back to expert-system development. A whole scientific discipline called knowledge engineering is devoted to the task of providing tools and methods for supporting the knowledge acquisition and formalization process [Studer et al., 1998]. In connection with the interest in ontologies as a key technology in knowledge and information sharing, the term "ontological engineering" has become popular [Farquhar and Gruninger, 1997] and a number of methodologies for creating ontologies have been proposed [Gomez-Perez and Juristo, 1997, Uschold, 1996]. However, these methods are very general in nature as they aim at providing general guidelines for all kinds of ontologies and purposes. We therefore propose a specialized strategy for the explication of information semantics.

In this chapter, we first review existing work on ontology engineering. We review existing methodologies and focus on approaches that have been proposed in combination with the task of integrating and sharing information. We conclude that existing methodologies do not address the problem of building a representational infrastructure like the one introduced in Chap. 2. We present an iterative approach for building source ontologies and shared vocabularies in a bottom-up fashion. We discuss the general process and useful resources and illustrate the method using a real life integration task.

4.1 Ontological engineering

The previous sections provided information about the use and importance of ontologies. Hence, it is crucial to support the development process of ontologies. In this section, we will describe how the systems provide support for the ontological engineering process.

Recently, several articles about ontological developments have been published. Jones and others [Jones et al., 1998] provide a short overview of existing approaches (e.g. METHONTOLOGY [Gomez-Perez and Juristo, 1997] or TOVE [Fox and Grninger, 1998]). Uschold [Uschold, 1996] and Gomez-Perez and others [Gomez-Perez et al., 1996] propose methods with phases that are independent of the domain of the ontology. These methods are of a good standard and can be used for comparisons. In this section, we focus on the proposed method from Uschold and Gruninger as a “thread” and discuss how the integrated systems evaluated in this chapter are related to this approach.

Uschold defines four main phases:

1. Identifying a purpose and scope: specialization, intended use, scenarios, set of terms including characteristics and granularity
2. Building the ontology:
   (a) Ontology capture: knowledge acquisition, a phase interacting with requirements of phase 1.
   (b) Ontology coding: structuring of the domain knowledge in a conceptual model.
   (c) Integrating existing ontologies: re-use of existing ontologies to speed up the development process of ontologies in the future.
3. Evaluation: verification and validation.
4. Guidelines for each phase.

In the following paragraphs we describe integration systems and their methods for building an ontology. Further, we discuss systems without an explicit method where the user is only provided with information in the direction in question. The second type of systems can be distinguished from others without any information about a methodology. This is due to the fact that they assume that ontologies already exist.

Infosleuth

This system semi-automatically constructs ontologies from textual databases [Hwang, 1999]. The methodology is as follows: first, human experts provide a small number of *seed words* to represent high-level concepts. This can be seen as the identification of purpose and scope (phase 1). The system then processes the incoming documents, extracting phrases that involve seed words, generates corresponding concept terms, and then classifies them into the ontology. This can be seen as ontology capturing and part of coding (phases 2 and 2). During this process the system also collects seed word-candidates for the next round of processing. This iteration can be completed for a predefined number of rounds. A human expert verifies the classification after each round (phase 3). As more documents arrive, the ontology expands and the expert is confronted with the new concepts. This is a significant feature of this system. Hwang calls this "discover-and-alert" and indicates that this is a new feature of his methodology. This method is conceptually simple and allows effective implementation. Prototype implementations have also shown that the method works well. However, problems arise within the classification of concepts and distinguishing between concepts and non-concepts.

Infosleuth requires an expert for the evaluation process. When we consider that experts are rare and their time is costly this procedure is too expert-dependent. Furthermore, the integration of existing ontologies is not mentioned. However, an automatic verification of this model by a reasoner would be worthwhile considering.

KRAFT

This system offers two methods for building ontologies: the building of shared ontologies [Jones et al., 1998] and extracting of source ontologies [Pazzaglia and Embury, 1998].

The steps of the development of shared ontologies are (a) *ontology scoping*, (b) *domain analysis*, (c) *ontology formalization*, (d) *top-level ontology.* The minimal scope is a set of terms that is necessary to support the communication within the KRAFT network. The domain analysis is based on the idea that changes within ontologies are inevitable and the means to handle changes should be provided. The authors pursue a domain-led strategy [Patil et al., 1991], where the shared ontology fully characterizes the area of knowledge in which the problem is situated. Within the ontology formalization phase the fully characterized knowledge is defined formally in classes, relations and functions. The top-level ontology is needed to introduce predefined terms/primitives.

If we compare this to the method of Uschold and Gruninger we can conclude that ontology scoping is weakly linked to phase 1. It appears that ontology scoping is a set of terms fundamental for the communication within the network and therefore can be seen as a vocabulary. On the other hand, the authors say that this is a *minimal* set of terms, which implies that more terms exist. The domain analysis refers to phases 1 and 2, whereas the ontology formalization refers to phase 2. Existing ontologies are not considered.

Pazzaglia and Embury [Pazzaglia and Embury, 1998] introduce a bottom-up approach to extract an ontology from existing shared ontologies. This extraction process consists of two steps. The first step is a syntactic translation from the KRAFT exportable view (in a native language) of the resource into the KRAFT schema. The second step is the ontological upgrade, a semi-automatic translation plus knowledge-based enhancement, where local ontology adds knowledge and further relationships between the entities in the translated schema.

This approach can be compared to phase 2, the integration of existing ontologies. In general, the KRAFT methodology lacks the evaluation of ontologies and the general-purpose scope.

Most Information integration systems, such as PICSEL, OBSERVER, BUSTER and COIN either have no methods or do not discuss them to create ontologies. After reading papers about these various systems it becomes obvious that there is a lack of a “real” methodology for the development of ontologies. We believe that the systematic development of the ontology is extremely important and therefore the tools supporting this process become even more significant.

4.2 Building an ontology infrastructure for Information sharing

The integration process sketched above relies on the existence of a shared ontology suitable to define concepts from all terminologies to be integrated in sufficient detail. This requirement is a challenge with respect to ontology building. In order to support this difficult task, we propose a development strategy that is tailored to the purpose of building shared ontologies. In this section we give an overview of the development process.

The process

The proposed strategy is based on stepwise refinement. It consists of five steps that are executed in an iterative process resulting in a partial specification of the shared ontology. The last step of each run is an evaluation step that triggers one of the previous steps in order to extend and refine the ontology if necessary. Fig. 4.1 illustrates the process model; the individual steps are briefly described below.

1. *Finding common concepts.* The first step is to examine the translation task. Asking the question “what do I want to translate?” leads to a concept that subsumes all classes from the source and destination systems. Because this concept makes a semantic translation from one source into another

possible we call it a bridge concept. By defining its properties and attribute values we achieve the needed shared vocabulary. The most general bridge concept is “top”, a concept that subsumes every other possible concept. For an exact classification it is recommended to choose the bridge concept as concretely as possible. If needed, more than one bridge concept can be defined to enable semantic translation.

2. *Definition of properties.* The next step is to define properties that describe the chosen bridge concepts. A car, for instance, can be described through its color, its brand, its price, etc.
3. *Finding property values.* Once we have defined the properties, we search for values which can fill the attributes. These “fillers” are the main part of the shared vocabulary.
4. *Adapt ontology.* The use of existing sources of information will not always be sufficient to describe all concepts of an information source. We sometimes have to handle very specific distinctions made in a system that hardly occur in standard terminologies. In order to capture these subtle differences we have to invent application-specific terms as part of the shared vocabulary.
5. *Refine definitions.* The introduced strategy follows the “evolving” life cycle. It allows the engineer to step back all the time to modify, add and remove ontology definitions, e.g. refining the bridge concept or integrating further taxonomies into the shared vocabulary.

Each of the steps modifies a different aspect of the shared ontology. While step 1 is concerned with the central concept definition, step 2 defines slots, step 3 integrates existing taxonomies and step 4 generates application-specific taxonomies. These facts are useful in order to determine where to go back to if the evaluation step reveals the inability to describe a certain aspect of a terminology to be integrated.

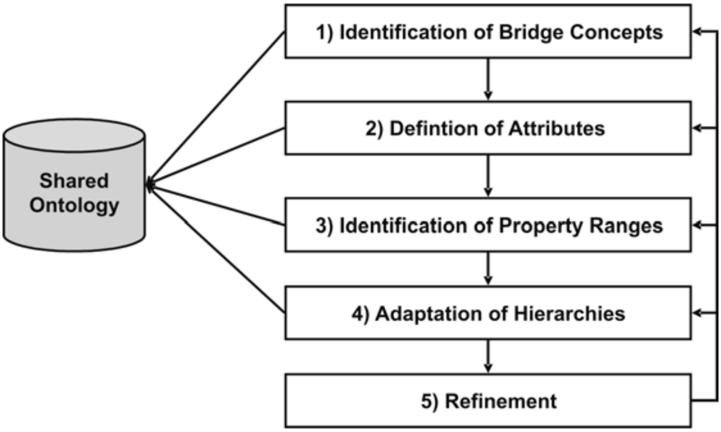
Shared Ontology
1) Identification of Bridge Concepts
2) Defintion of Attributes
3) Identification of Property Ranges
4) Adaptation of Hierarchies
5) Refinement

Fig. 4.1. Steps of the development process

Sources of information

The use of the ontology to be built as a common basis for communication between systems makes it necessary to stay as closely as possible to a vocabulary and conceptualization of the domain that is widely accepted as a standard. In order to meet this requirement, we use several sources of information to build upon. These information sources are existing ontologies and thesauri as well as scientific classifications and data catalogues.

- Top-level ontologies are mainly used to find the bridge concept which acts as a template for the definition of all terms to be translated. In most cases, the bridge concept is obvious; however, the use of an upper level ontology provides us with a vocabulary which is partly standardized.
- Scientific classifications are another form of standards describing the conceptualization of a domain. Classifications like taxonomies of animals or plants are common knowledge which can be used to specify concepts from domain-specific ontologies.
- Domain thesauri contain typical terms used in an application domain; therefore, they are a natural source for finding concept names for the shared ontology. Further, many thesauri contain at least free-text definitions of the terms included. These definitions provide guidance for the definition of concepts.
- Linguistic thesauri are used to supplement information taken from domain-specific thesauri. In contrast to the specialized vocabulary defined in domain-specific thesauri, linguistic thesauri can be used to identify correspondences between terms found in different information sources. Especially, we use linguistic thesauri to expand the search for definitions of terms to their synonyms.
- Data catalogues finally contain the definitions of the terminology to be modelled. Therefore, they define the concepts to be modelled and are the basis for evaluating the expressiveness of the shared ontology at a specific point in the modelling process.

In the course of the modelling process, we stick as closely as possible to the information from the sources mentioned above. Therefore, the selection of these sources, though not discussed in this book, is an important step when building a shared ontology.

4.3 Applying the approach

We performed a case study in order to assess the general strategy described above. In the following we will describe the task of this case study and give an impression of how the strategy helps to build the models needed to solve it.

4.3.1 The task to be solved

Geographical information systems normally distinguish different types of spatial objects. Different standards exist specifying these object types. These standards are also called catalogues. Since there is more than one standard, these catalogues compete with each other. To date, no satisfactory solution has been found to integrate these catalogues. In our evaluation we concentrate on different types of areas distinguished by the type of use.

In order to address the semantic translation problem we assume a scenario where the existing land-administration database that is normally based on the ATKIS catalogue, which is the official standard for most administrations, should be updated with new information extracted from satellite images of some area. Satellite images are normally analyzed using image-processing techniques resulting in a segmentation of different areas which are classified according to the CORINE landcover nomenclature, a standard for the segmentation and classification of satellite images. The process of updating the land-administration system with this new data faces two main problems:

1. The boundaries of the objects in the database might differ from the boundaries determined on the satellite image.
2. The class information attached to areas on the satellite image and the type information in the land-administration system do not match.

The first problem is clearly beyond the scope of our investigation, but the second is a perfect example of a semantic translation problem. A successful integration of the two information sources will come with the following benefits for the user of the systems: (a) *integrated views* and (b) *verification.* An integrated view from the user's perspective merges the data between the catalogues. This process can be seen as two layers which lay on top of each other. The second option gives users the opportunity to verify ATKIS-OK-250 data with CORINE landcover data or vice versa.

The basis for our experiment is a small CORINE landcover dataset containing information about the town “Bad Nenndorf” in Lower Saxony. This dataset is available from the German Environmental Agency in different formats and classifications and can therefore be used to compare and evaluate results. In our case study, we want to find out whether land-use classes from the CORINE landcover dataset can be semantically translated into the classification used by the ATKIS catalogue. Such a translation could be the basis for both the generation of an integrated view of the information in both systems and for a validation of ATKIS data with up-to-date satellite images. Fig. 4.3.1 illustrates the integration problem.

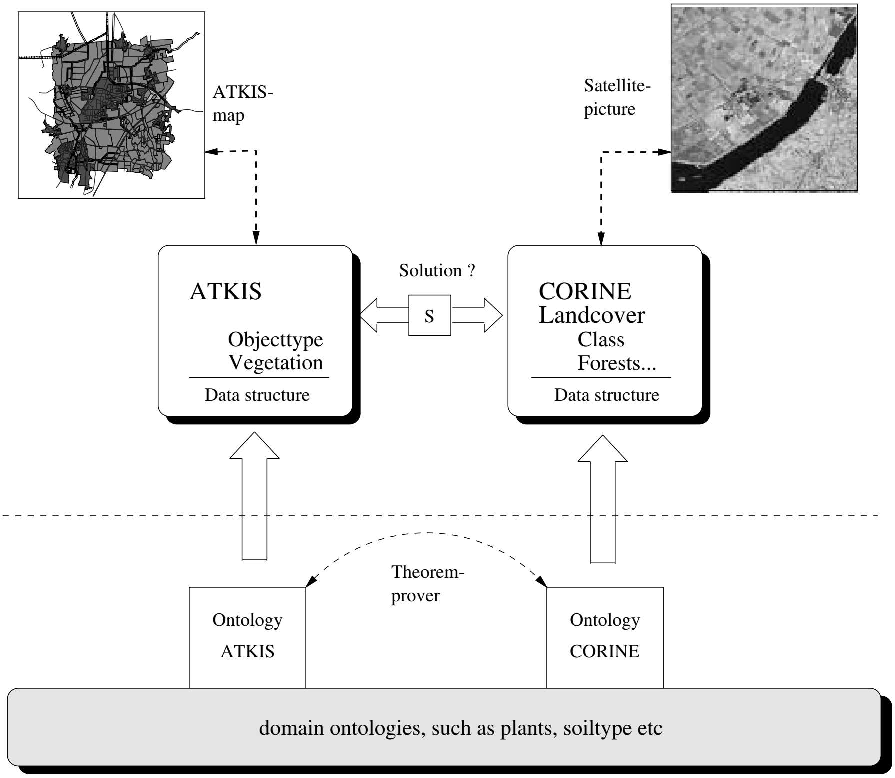

Fig. 4.2. Deductive integration of geographic information

4.3.2 The Information Sources

The ATKIS catalogue [AdV, 1998] is an official information system in Germany. It is a project of the head surveying offices of all the German states. The working group offers digital landscape models with different scales from 1:25000 up to 1:1000000 with a detailed documentation in corresponding object catalogues. We use the large-scale catalogue OK-1000. This catalogue offers several types of objects including definitions of different types of areas. Fig. 4.3 shows the different types of areas defined in the catalogue.

CORINE landcover [European Environmental Agency, 1999a] is a result of the CORINE programme the European Commission carried out from 1985 to 1990. The results are essentially of three types, corresponding to the three aims of the programme: (a) an information system on the state of the environment in the European Community has been created (the CORINE system). It is composed of a series of databases describing the environment in the European Community, as well as of databases with background information. (b) Nomenclatures and methodologies were developed for carrying out the programme, which are now used as the reference at the Community level. (c) A systematic effort was made to coordinate activities with all the bodies involved in the production of environmental information especially at international level. The nomenclature developed in the CORINE programme can be seen as another catalogue, because it also defines a taxonomy of area types (see Fig. 4.4) with a description of characteristic properties of the different land types.

The task of this example is that the data of the CORINE database has to be converted into the ATKIS database. Of course, this transformation can be viewed as a special case of an integration task demonstrating all the problems which can occur. Besides the obvious structural heterogeneity problems, the main problem lies in the reconciliation of the semantic heterogeneity caused by the use of different classification schemes.

The classification schemes of land-use types in Figs. 4.3 and 4.4 illustrate this problem. The set of land types chosen for these catalogues are biased by their intended use: while the ATKIS catalogue is used to administrate human activities and their impact on land use in terms of buildings and other installations, the focus of the CORINE catalogues is on the state of the environment in terms of vegetation forms. Consequently, the ATKIS catalogue contains fine-grained distinctions between different types of areas used for human activities (i.e. different types of areas used for traffic and transportation) while natural areas are only distinguished very roughly. The CORINE taxonomy on the other hand contains many different kinds of natural areas (i.e. different types of cultivated areas) which are not further distinguished in the ATKIS catalogue. On the other hand, areas used for commerce and traffic are summarized in one type.

Despite these differences in the conception of the catalogues the definition of the land-use types can be reduced to some fundamental properties. We identified six properties used to define the classes in the two catalogues. Beside *size* and *general type of use* (e.g. production, transportation or cultivation) the *kinds of structures* built on top of an area, the *shape of the ground* and *natural vegetation* as well as kinds of *cultivated plants* are discriminating characteristics.

4.3.3 Sources of knowledge

For this specific integration task we chose several sources of information to be used for guiding the development process. We briefly describe these sources in the following.

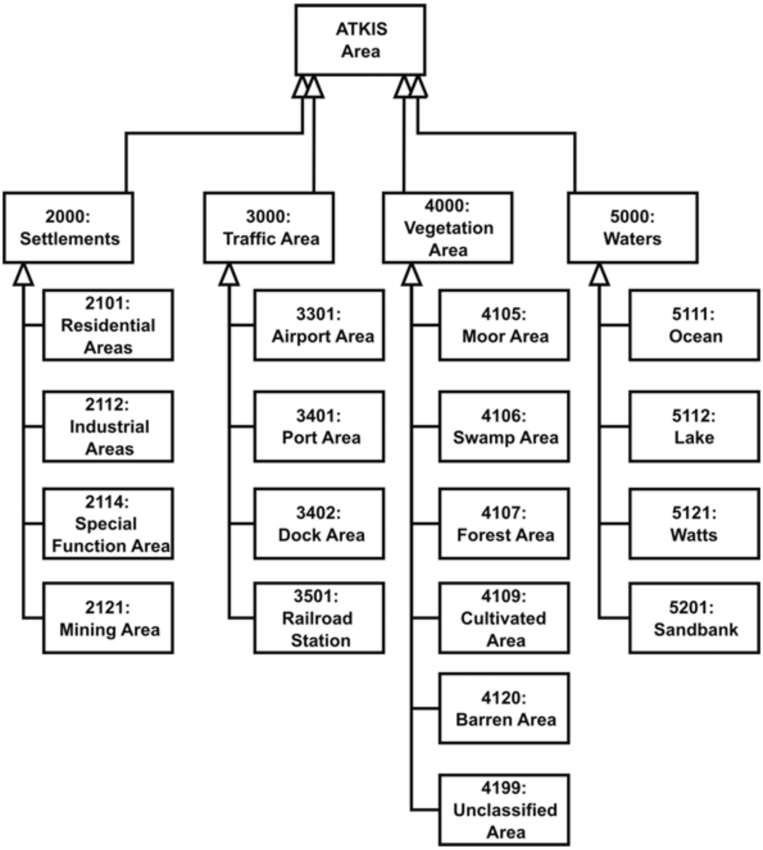

Fig. 4.3. Taxonomy of land-use types in the ATKIS OK-1000 catalogue

UpperCyc ontology

Upper Cyc, developed by the CyCorp corporation [Lenat, 1995] (http://www.cyc.com), is an upper-level ontology that captures approximately 3 000 terms of the most general concepts of human consensus reality. There is also a full Cyc knowledge base (KB) including a vast structure of more specific concepts descending below Upper Cyc, the so-called top-level ontology. It contains millions of logical axioms – rules and other assertions – which specify constraints on the individual objects and classes found in the real world. Therefore the Upper Cyc ontology provides a sufficient common ground for applications. We chose Cyc as a reference for selecting the bridge concept, because it provide a large number of higher level concepts.

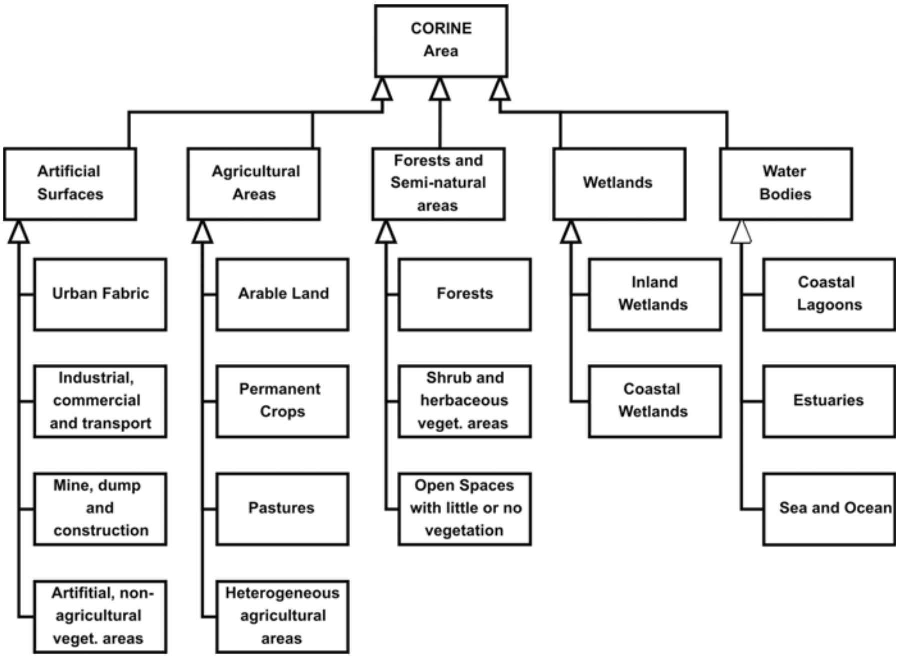
CORINE Area
Artificial Surfaces
Agricultural Areas
Forests and Semi-natural areas
Wetlands
Water Bodies
Urban Fabric
Industrial, commercial and transport
Mine, dump and construction
Artifitial, non-agricultural veget. areas
Arable Land
Permanent Crops
Pastures
Heterogeneous agricultural areas
Forests
Shrub and herbaceous veget. areas
Open Spaces with little or no vegetation
Inland Wetlands
Coastal Wetlands
Coastal Lagoons
Estuaries
Sea and Ocean

Fig. 4.4. Taxonomy of land-use types in the CORINE landcover nomenclature

GEMET

The general multilingual environmental thesaurus GEMET [European Environmental Agency, 1999b] is a poly-hierarchically structured thesaurus which covers approximately 5 400 terms and their definitions organized by groups, themes and terms. GEMET has been created by merging different national and international thesauri. Analysis and evaluation work of numerous international experts and organizations led to a core terminology of generalized environmental terms and definitions. GEMET ensures validated indexing and cataloguing of environmental information all over Europe. Where available, synonyms or alternate terms can be found likewise. We chose the GEMET thesaurus as a source for definitions of concepts and to supplement the information obtained from Cyc with domain-specific information. These definitions provide for example insight into useful properties of classes.

WordNet

WordNet [Fellbaum, 1998], developed by the Cognitive Science Laboratory at Princeton University, is an on-line lexical reference system whose design is inspired by current psycholinguistic theories of human lexical memory. English nouns, verbs, adjectives and adverbs are organized into synonym sets, each representing one underlying lexical concept. Different relations link the synonym sets. WordNet was mainly used as a source of synonymy information needed to look up concepts across the other knowledge sources (e.g. to find the equivalent of a concept from Cyc to look up the domain-specific defintion).

Standard taxonomies

Scientific taxonomies can be found in many sources, like books or the Internet. For this example we looked at the Google Webdirectory (http://directory.google.com/Top/Science/Biology/Flora_and_Fauna) to obtain a classification of plant life. It is in no circumstances complete, but it satisfies our needs in this case study. We chose the classification of plants to determine possible fillers for the properties of a class as many land types are mostly defined by the vegetation found (e.g. mixed forest).

4.4 An example walkthrough

Based on the information described above we built up a first version of a shared ontology which should be used to solve the integration task mentioned in the last section. In this section we sketch the first development cycle of this ontology using concrete modelling activities to illustrate the different steps of our strategy using modelling example from the CORINE classification. The corresponding definitions of ATKIS concepts that will also be created in the different steps discussed below are not shown.

Step 1: finding bridge concepts

Looking at the given example scenario as described in Sect. 4.2 it is quite obvious to choose a concept like “area” or “region”, because all land-use classes are some kind of special “regions” or, in other words, “region” subsumes all land-use classes. We search for the term “region” in the Upper CYC and get the following definition:

> *“GeographicalRegion: a collection of spatial regions that include some piece of the surface of PlanetEarth. Each element of GeographicalRegion is a PartiallyTangible entity that may be represented on a map of the Earth. This includes both purely topographical regions like mountains and underwater spaces, and those defined by demographics, e.g. countries and cities* [· · ·]*”.*

Fig. 4.5 shows the hierarchical classification of the concept in the Upper Cyc. The definition fits very well, so finally we choose ”Geographical Region” as our bridge concept. For further refinement we write it down in the OWL notation.

```
Class(Geographical-Region)
```

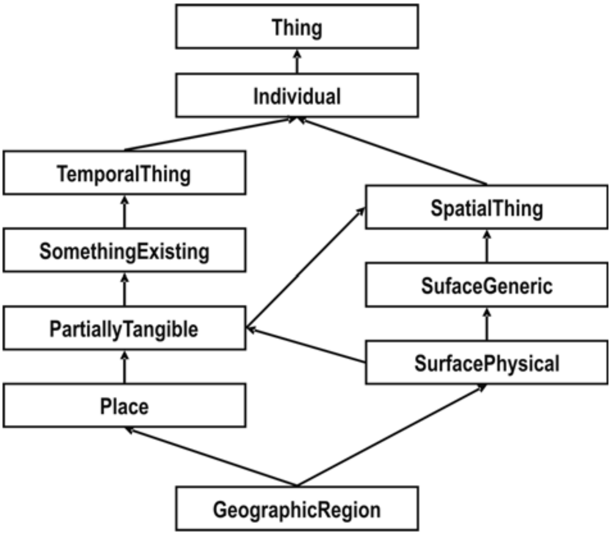

Fig. 4.5. Geographical region in Upper Cyc

Step 2: definition of properties

Now we have to find possible attributes for the bridge concept. We look for "Geographical Region" in GEMET, but the search does not give any results. In that case the decomposition of the search phrase may give better results. For "Geography" and "Region" we get these definitions out of GEMET:

> *"Geography: The study of the natural features of the Earth's surface, comprising topography, climate, soil, vegetation, etc and Man's response to them."*

> *"Region: A designated area or an administrative division of a city, county or larger geographical territory that is formulated according to some biological, political, economic or demographic criteria."*

In the definition of "geography", some attributes are clearly recognizable. For example, climate, soil, vegetation and human activities. We use vegetation to illustrate the next steps in our method. Vegetation is a biological criterion that defines a region, and it is also part of the scientific field "geography". We update the bridge concept by defining a slot "vegetation" and adding it to the bridge concept.

```
 Class(Geographical-Region)
ObjectProperty(vegetation domain(Geographical-Region))
```

Step 3: integration of standard taxonomies

To get possible "attribute values" or "fillers" for the slot "vegetation", we take another look at GEMET. Vegetation is defined as:

> *"The plants of an area considered in general or as communities* [· · ·]*; the total plant cover in a particular area or on the Earth as a whole."*

We also check the synonym "flora", found in WordNet:

> *"The plant life characterizing a specific geographic region or environment."*

The attribute "vegetation" or "flora", can be filled with terms out of plant life like "tree" or "rose" for instance. A good top concept is "plants", because many scientific taxonomies of plants exists. The Swedish botanist Carlous Linaeus established in 1753 a classification of plants. His work is considered the foundation of modern botanical nomenclature. In the Google Webdirectory we can access the plant kingdom with more than 10 000 entries on-line. We integrate this taxonomy into our vocabulary, because we need concept from it to distinguish concepts in our information sources through the reference to this hierarchy.

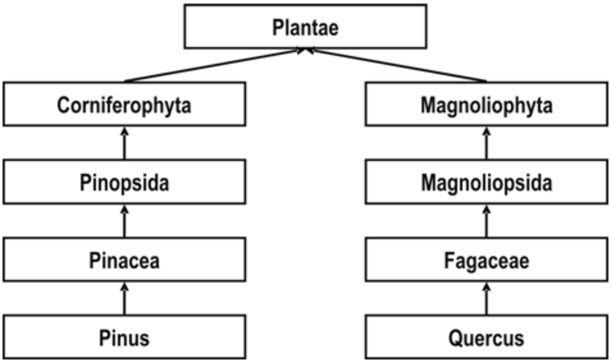

Fig. 4.6. Extract from scientific plant taxonomy

Now it is possible to describe classes from the land-use catalogues. The term "coniferous forest" in the CORINE context is defined as:

> *"Vegetation formation composed principally of trees, including shrub and bush understories, where coniferous species predominate."*

In our vocabulary we find the term "coniferophyta", comprising the conifers, which are trees or shrubs that bear their seeds in cones, without the protection of a fruit, like angiosperms. This leads to the following OWL class:

```
    SubClassOf(Coniferous-Forest interSectionOf(
                Geographical-region
                restriction(vegetation
                    allValuesFrom(Coniferophyta))))
```

The division *magnoliophyta* of the plant kingdom consists of those organisms commonly called the flowering plants, or angiosperms. The flowering plants are the source of all agricultural crops, cereal grains and grasses, garden and road-side weeds, familiar broad-leaved shrubs and trees and most ornamentals. So, it is easy to describe the next CORINE class "broad-leaved forest":

```
   SubClassOf(Broad-leaved_Forest intersectionOf(
                Geographical-region
                restriction(vegetation
                    allValuesFrom(Magnoliophyta))))
```

A "mixed forest" in the CORINE nomenclature consists of conifers and broad-leaved trees.

```
   SubClassOf(Mixed_Forest intersectionOf(
                Geographical-region
                restriction(vegetation
                    someValuesFrom(Magnoliophyta))
                restriction(vegetation
                    someValuesFrom(Coniferophyta))))
```

Step 4: adapt vocabulary

A closer look at the definition of the CORINE forest classes reveals that the classes are defined through the existence of trees and shrubs. Just using the term "magnoliophyta" does not prevent the classification of a region covered with orchids as a broad-leaved forest (orchidaceae is a subclass of magnoliophyta). The mentioned taxonomy classifies plants according to their way of reproduction, therefore distinguishing angiosperm and gymnosperm trees, shrubs and flowers. To handle this problem we need a more general distinction.

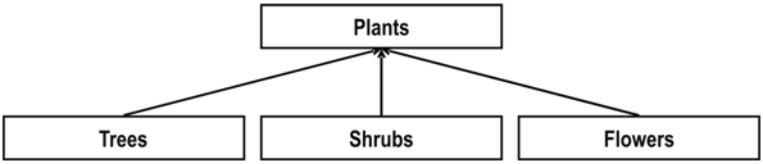

Fig. 4.7. Supplementary plant classification

Fig. 4.7 shows a simple extension of the vocabulary that enables a more robust definition of the CORINE forest classes.

```
SubClassOf(Coniferous-Forest intersectionOf(
            Geographical-region
            restriction(vegetation
                allValuesFrom(Coniferophyta))
            restriction(vegetation
                allValuesFrom(unionOf(trees shrubs)))))

SubClassOf(Broad-leaved_Forest intersectionOf(
            Geographical-region
            restriction(vegetation
                allValuesFrom(Magnoliophyta))
            restriction(vegetation
                allValuesFrom(unionOf(trees shrubs)))))

SubClassOf(Mixed_Forest intersectionOf(
            Geographical-region
            restriction(vegetation
                someValuesFrom(Magnoliophyta))
            restriction(vegetation
                someValuesFrom(Coniferophyta))
            restriction(vegetation
                allValuesFrom(unionOf(trees shrubs)))))
```

The shared vocabulary developed so far allows us to specify many different vegetation areas found in the land-use catalogues:

```
SubClassOf(Pastures intersectionOf(
            Geographical-region
            restriction(vegetation allValuesFrom(Poaceae))))

SubClassOf(vineyards intersectionOf(
            Geographical-region
            restriction(vegetation allValuesFrom(Vitis))))

SubClassOf(Rice_fields intersectionOf(
            Geographical-region
            restriction(vegetation allValuesFrom(Oryza))))
```

This definition might seem to be too restrictive, because it does not allow any other plants other than the dominant species. Our goal here is not, hewever, to provide a complete description of the concepts in terms of all the vegetation that might be found. Such a modelling approach would be much to big an effort to make sense. We rather want to characterize a concept by the properties that distinguishes it from the other concepts in the hierarchy. The definitions above satisfy this requirements. In order to make this more explicit the vegetation property can be read as "dominant vegetation form".

Step 5: evaluation and revision

Not all CORINE landcover classes can be described after this first process cycle. “Mineral extraction sites”, for instance, are defined as:

> *“Areas with open-pit extraction of minerals (sandpits, quarries) or other minerals (opencast mines). Includes flooded gravel pits, except for river-bed extraction.”*

No vegetation is mentioned, so the bridge concept must be refined. We go back to step 2 “defining properties” and search for another attribute. The definitions of “region” and “geography” show some anthropological aspects, like “Man’s response” or economic criteria. So we define a new slot ’anthroposphere’ and add it to our bridge concept:

```
Class(Geographical-Region)
ObjectProperty(vegetation
                domain(Geographical-Region))
ObjectProperty(anthroposphere
                domain(Geographical-Region))
```

In the topic area “anthroposphere” of the GEMET thesaurus we find the term “mining district”, a district where mineral exploitation is performed. We integrate the partial taxonomy into the vocabulary (Fig. 4.8).

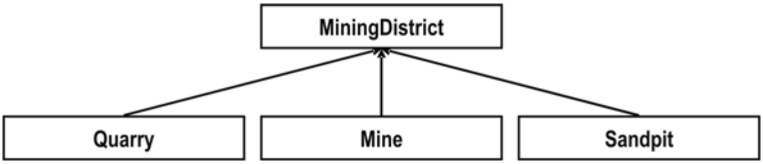

Fig. 4.8. Mining sites from the GEMET thesaurus

This special vocabulary can be used to simulate one-to-one mappings by using equality axioms. The CORINE class “mineral extraction sites” could be described as follows.

```
SubClassOf(Mineral-extraction-sites intersectionOf(
            Geographical-region
            restriction(anthroposphere
                allValuesFrom(mining-district)))))
```

In a similar way, we proceed by iterating the process cycle until all terms from the two catalogue systems can be modelled as a specialization of the bridge concept. A further advantage of this strategy is the fact that the same process will be employed when additional terminologies are to be integrated as well. We cannot guarantee that the shared ontology also covers a new terminology, but our strategy already provides guidance for the adaption of the ontology.

4.5 Conclusions

In real applications the most important question is often not how to arrange ontologies, but how to actually build these ontologies. This problem has been widely recognized and some methodologies have been developed to support the development of ontologies. In most cases, these methodologies are very general and only provide basic guidance for the development of an ontology infrastructure. In our approach the notion of a shared vocabulary is essential and the development of this vocabulary therefore deserves special attention. We had good experiences with a strategy that follows a bottom-up approach that takes the actual integration problem as a starting point and consults general models like top-level ontologies and linguistic resources only if necessary. The resulting vocabularies are general enough to cover at least a certain class of integration problems. We think that this is more valuable than a general top-down approach because it solves real world problems without losing the connection to basic ontological principles.

The examples given above already show that the method leads to better results than an early hands-on approach described in [Stuckenschmidt et al., 2000]. In this early case study, we developed the shared vocabulary solely by relying on textual description of the two catalogues mentioned above. The development strategy proposed here results in a shared model that uses mostly standardized terms and is well integrated with existing higher-level ontologies.

We also managed to describe more concepts with fewer properties. The use of the vegetation property for example turned out to be sufficient for describing about half of all concepts from both information sources. We explain this with the richer vocabularies for describing different vegetation types we got from scientific classifications.

An interesting side effect of the more controlled development is a harmonization of the structure of logical expressions used to define concepts. We explain this by the fact that the strategy forces us not to describe a concept completely without comparing it to other definitions. The strategy rather forces us to define restrictions for a particular property for many concepts in parallel. This direct comparison makes it easier to capture the specific structure of the logical expression required in contrast to the definition of other concepts.

Further reading

Further information about the information sources used for ontology development can be found in the official documentation published, by the German administration [AdV, 1998] and the European environment agency [European Environmental Agency, 1999a]. The Cyc ontology, the WordNet lexical database and the GEMET thesaurus used for identifying and characterizing the bridge concept are described in [Lenat, 1995], [Fellbaum, 1998] and [European Environmental Agency, 1999b], respectively. A detailed documentation of an earlier attempt to model the information sources with the ontology language OIL is described in [Stuckenschmidt et al., 2000].