(Draft) Proceedings of the 15th Database and Experts Systems Applications - DEXA'04 (1st International Workshop on Geographic Information Management - GIM'04). 2004, p. 878-882.

Management of nested collections of resources in Spatial Data Infrastructures

J. Nogueras-Iso, F.J. Zarazaga-Soria, P.R. Muro-Medrano
Computer Science and Systems Engineering Department,University of Zaragoza
María de Luna, 1. 50018-Zaragoza (Spain)
jnog, javy, prmuro@unizar.es
http://iaaa.cps.unizar.es

Abstract

The management of collections of resources is an important issue in Digital Libraries. The modelling of collections provides added-value because they facilitate the organization of resources and special services may be tailored to the characteristics of the collection. This paper will provide a metadata solution to manage nested collections in digital library catalogs, which is based on XML technologies and concepts derived from knowledge bases. The management of collections is particularly relevant in the context of Geolibraries and Spatial Data Infrastructures. Given the high volumes of geographic information, it is very frequent to find collections that arise as a result of the fragmentation of geographic resources into datasets of manageable size and similar scale. However, the concepts presented here are extensible to any type of Digital Library collections.

1. Introduction

As regards the cataloguing of geographic resources, an important circumstance to take into account is the existence of collections or aggregation of geographic resources (or datasets) that can be considered as a unique entity. Most of these collections arise as a result of the fragmentation of geographic resources into datasets of manageable size and similar scale. In this sense, for example, the *Spanish National Geographic Institute* (*IGN*) offers distinct versions of its products (*Cartographic Numeric Base BCN*, *National Topographic Map MTN*, *Digital Terrain Model MDT*,...) according to different scales: *BCN200* identifies the *BCN* at 1:200,000 scale; *BCN25* identifies *BCN* at 1:25,000 scale and so on. Each product-version pair compiles the set of files into which the Spanish territory was divided so as to provide, at the scale required, a number of files with reasonable size. Those files are usually named "tiles" and *IGN* establishes for each scale the numbering and spatial extent covered by these tiles. Besides, each aforementioned product may be in turn composed of several information layers. For example, each *BCN* tile is composed of the following thematic layers: administrative divisions; altimetry; hydrography and coasts; buildings and constructions; communication networks; utilities; and geodetic vertexes. That is to say, it is also common to organize resources in more than one level of aggregation, originating nested collections. By nested collections it is meant that a collection can be included as a part of another collection. This recursive definition of collections enables the hierarchical organization of resources in a repository.

When providers or distributors of geographic information want to publish the content of their holdings, they must provide standardized descriptions of their datasets (metadata), which are later incorporated into data catalogs and clearinghouses. The creation and maintenance of geographic metadata is a time consuming and thorough process. This circumstance is especially problematic if a collection of thousands of datasets must be documented. On one hand, the datasets belonging to the same collection share a high percentage of meta-information that must be replicated multiple times. And on the other hand, users of geographic information are accustomed to manage the entire collection as a unique entity (e.g. the *National Topographic Map* at scale 1:50,000), which should be returned by data catalogs as a unique result instead of displaying the complete list of thousands of files that conform the collection.

The problem of how to describe collections within metadata is an important issue in new proposals for geographic information metadata standards (e.g., ISO19115 [7] or Remote extensions of CSDGM [2]). Thus, most of these metadata standards define elements to point at related resources, usually by means of a string or number conforming to a formal identification system. However, a catalog system can not manage collections just enabling librarians to manually edit the fields concerned with these links. There are several aspects that justify a more complex implementation of collections. Firstly, the resources (and metadata records describing them) must be uniquely identified, at least within the local catalog. Thus, all the references among the aggre-

(Draft) Proceedings of the 15th Database and Experts Systems Applications - DEXA'04 (1st International Workshop on Geographic Information Management - GIM'04). 2004, p. 878-882.

gate and the parts must be always up-to-date whenever a component of the aggregation is added or removed. Secondly, the components that form part of a collection usually share a high percentage of meta-information (e.g., abstract, topic category, etc.). There are metadata elements whose content could be inherited from the metadata record that describes the collection. But if the catalog does not provide an automatic mechanism to inherit meta-information, metadata creators must replicate common descriptions for each dataset. And thirdly, some values of the metadata elements (e.g. the temporal or spatial extent) in the collection metadata record are aggregated or averaged over the values of the components of the collection.

The objective of this paper will be to provide a metadata solution to manage nested collections of geographic resources, which is based on XML technologies and concepts derived from knowledge bases. The most accepted way to exchange metadata is by means of XML documents, whose syntax is enforced by control files in the form of DTDs or XML-Schemas. Thus, a system managing metadata records as XML documents will be highly independent of the structure of metadata standards. This paper proposes the construction of catalog services over a knowledge base component, which is able to store the different types of metadata schemas supported, the aggregation relations established among these schemas, and the inference mechanisms that these relations will provide.

2. Related work

According to [6], the precedent of the management of collections in digital libraries can be found in the world of online bibliographic services, which sum up the content of materially significant databases. On the other hand, as traditional libraries gave public access to their catalogs via the Internet, several standardization initiatives appeared to describe the contents of a collection such as: the Encoded Archival Description standard for the encoding archival finding aids to collections of materials; the Z39.50 Profile for access to digital collections; or the Resarch Support Libraries Programme Collection Description Project [9].

One of the most relevant works to facilitate the access to digital library collections is the STARTS protocol [4]. This protocol for internet search and retrieval facilitates the task of querying multiple document sources, namely text collections accessed via search engines. The goal of STARTS is that the search engines implementing the protocol will assist a meta-searcher in choosing the best sources to evaluate a query, evaluating the query at these resources, and merging the query results from these sources. The basis for the implementation of STARTS protocol is the availability of *source metadata*, which describes the contents of the collection. This metadata consists of two pieces: the *source metadata attributes*, which includes information that a meta-searcher can use to rewrite the queries sent to the source as well as other attributes manually generated (e.g., abstract, contact or access constraints); and the *source content summary* which contains the information that is automatically generated such as the list of words that appear in the source, the statistics for each word listed, or the total number of documents in the source.

Within the context of geolibraries, a good example of a system handing collections is the Alexandria Digital Library (ADL) project [6]. Collections of geographically referenced items (maps, aerial photographs, satellite images, etc.) are described by means of: *collection level metadata*, which is a standardized description about the collection; and *item level metadata*, which are the individual descriptions of the items that form part of the collection. Similar to STARTS *source metadata*, *collection level metadata* is also divided into: contextual metadata (equivalent to STARTS *source metadata attributes*) and inherent metadata (the STARTS *source content summary*). The main contribution of ADL with respect to previous approaches is its geographic-oriented approach. Unlike text-oriented approaches (STARTS, bibliographic databases) it has identified the relevance of presenting graphic characteristics of the collections such as the visualizations of the geographic and temporal coverages.

3. The Metadata Knowledge Base

As mentioned in the introduction, the metadata records describing the components and the own collection as a whole only differ in a few set of metadata elements. Just observing the features of the most frequent types of collections, one could imagine the metadata elements that will probably differentiate the description of two components in the same collection. That is to say, instead of creating complete descriptions of each component in the collection manually, a system could automate this labor just having a high-level description of the entire collection and the specific values of just a few elements for each component. For instance, the *BCN200* product (mentioned in the introduction) is an example of a spatial collection (the components are spatially distributed to cover a wide area) that groups the files providing real data for each province in Spain. And the metadata describing each component uniquely differ in the specific *title* of the component; the *reference date*; the *geographic location identifier* (code and name of province); the *bounding box*; and the *coordinates reference system*. Besides, there are also some elements that are subject to be summarized and stored as common elements at collection level. For instance, for spatial collections, it results interesting to calculate the minimum *bounding box* that covers the *bounding boxes* of the components.

Therefore, our ideal catalog system should enable a flexible definition of metadata records (probably not con-

(Draft) Proceedings of the 15th Database and Experts Systems Applications - DEXA'04 (1st International Workshop on Geographic Information Management - GIM'04). 2004, p. 878-882.

strained to a specific metadata standard), provide inference mechanisms between relations established between metadata records, and, as mentioned in the introduction, support recursive levels of aggregations (i.e. nested collections). As a possible solution, such a catalog could be developed over a knowledge-base component. A Knowledge Base System is defined as a system that includes a knowledge base about a domain and programs that include rules for processing the knowledge and for solving problems relating to the domain. And one way to represent this knowledge could be based on the concept of ontology. An ontology is usually defined as an "explicit formal specification of a shared conceptualization" [5]. In the context of information systems and knowledge representation, the term ontology is used to denote a knowledge model, which represents a particular domain of interest. And more specifically in the context of metadata standards, the own structure of metadata standards (also called metadata schemas) can be considered as ontologies, where metadata records are the instances of those ontologies. Therefore, ontologies may be used to profile the metadata needs of a specific resource and its relationship with the metadata of other related resources. For instance, the ISO19115 geographic metadata standard [7] has been modelled as an ontology using the Protégé ontology editor (http://protege.stanford.edu/).

Fig. 1 shows the ontology describing the metadata needs for the collection and components of the *BCN200* using a frame-slot-facet representation [8]. There, each frame represents a different type of metadata schema. Although a metadata schema is usually structured in sections and subsections, for the sake of clarity, it is assumed that these schemas can be simplified into a flattened list of elements abstracting us from their complexity. The slots displayed inside frames correspond to some metadata elements of ISO19115. Besides, it can be observed that there are three types of relations between frames: the *is-a* hierarchy for creating more specific metadata schemas with more slots or modifying the slots of the parent frame; the *whole-part* hierarchy for establishing the relation between the metadata describing a collection and the metadata describing the components of that collection; and the *instance* hierarchy, which is used to relate instances of a metadata schema to the frame establishing its syntax.

Another question that may arise from the model in fig. 1 is why we should create two different schemas, *MD_Collection* and *MD_Component*, for the description of IGN products and components. In principle, all metadata instances should follow the syntax imposed by *MD_ISO19115*, which represents the ISO19115 standard. The answer to this question can be found in the different inference behavior of *MD_Collection* and *MD_Component* with respect to the *whole-part* relation. On one hand, the frame acting as *part* in a *whole-part* relation will obtain the value of a slot using one of the following prioritized ways: by means of the *part-if-needed* facet daemon, by using the own value of the slot, or by inheriting this value from the *whole*. The *part-if-needed* facet daemon returns a value obtained as the combination of slot values of the *part* and slot values of the *whole*. For instance, the *part-if-needed* of *datasetTitle* in fig. 1 concatenates the *datasetTitle* of the *whole* and the *geographicLocationIdentifier* of the *part*. On the other hand, a frame acting as *whole* will obtain the value of a slot using one of the following prioritized ways: by using the own value of the slot, or by means of the *whole-if-needed* facet daemon. This daemon is usually implemented as an aggregated function applied over the components of the aggregation. For instance, the *whole-if-needed* daemon of *geographicLocationBoundingBox* in fig. 1 computes the minimum bounding box covering the *geographicLocationBoundingBox* of the *parts*.

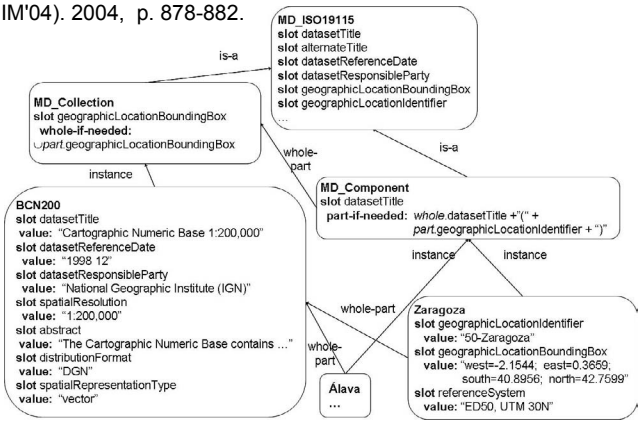

Figure 1. A frame-slot-facet representation

Although this frame-based solution seems to solve the problem of metadata duplication, the direct implementation by means of a frame-based language (understood in general terms as a knowledge-based approach) introduces important disadvantages. Firstly, the experience says that knowledge engineering specific tools have not been exploited enough in industrial applications [3]. A widely used ontology management tool like Protégé has not been tested with a real system containing more than 150,000 frames (classes & instances). However, a catalog managing collections could manage millions of metadata records. And secondly, using this frame-based solution, we need to define new frames not only for each metadata standard but also for each special behavior. The most accepted way to exchange metadata is by means of XML documents, whose syntax of this XML is enforced by control files in the form of DTDs or XML-Schemas. Given that standardization organizations usually publish these control files, the question is clear:

(Draft) Proceedings of the 15th Database and Experts Systems Applications - DEXA'04 (1st International Workshop on Geographic Information Management - GIM'04): 2004, p. 878-882.

"Why must we rewrite this syntax in the form of frames or other concept-based representations?". And thirdly, one of the desirable functionalities of the digital library catalog would be to provide collection statistics, which include histograms of spatial coverage or temporal coverage. However, frame-based languages do not usually provide many facilities for the work with complex data types.

Given these disadvantages, we have opted for our own implementation of knowledge bases that reinforces the role of relations and makes profit of XML technologies. On one hand, works like [1] encourage the improvement of semantics and inference mechanisms of *whole-part* relations in object-centered systems. In this case, our knowledge base enables the definition of *whole-part* relations where we have transferred the inference mechanisms previously found in the frames (*if-needed* facets). This way, frames are only focused in representing metadata, not in the behavior involved in *whole-part* relations. And on the other hand, the use of XML technologies increments the flexibility of the knowledge base. A knowledge base managing metadata records (instances) as XML documents and the syntax (frames) of those documents as XML-Schemas will be highly scalable and independent of the particular structure of each metadata standard.

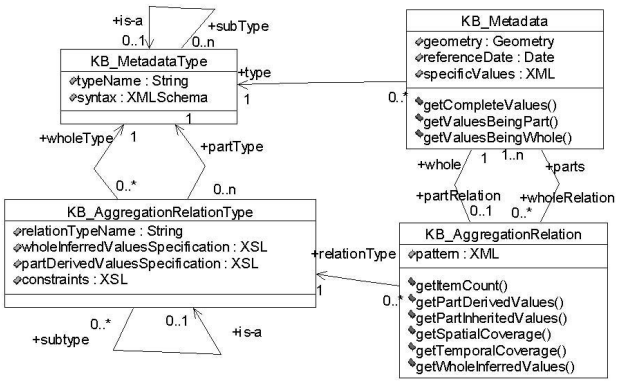

Figure 2. The Knowledge Base

Fig. 2 shows the object-oriented model for the implementation of this knowledge base, which uses a relational database (e.g., Oracle) as storage device and has been programmed in Java. As it can be observed, there are two differentiate parts in the model: on the left side, the classes that represent the metadata types and the relation types (the knowledge); and on the right side, the classes that represent the instances of these types.

Regarding the knowledege part of fig 2, the *KB_MetadataType* class represents the syntax of a metadata schema or standard. It has a *syntax* attribute which stores the XML-Schema that defines the syntax of a particular type of metadata. And the *KB_AggregationRelationType* class represents the types of relations established between two metadata types. The inference knowledge provided by the relation is specified in the attributes *wholeInferredValuesSpecification* and *partDerivedValuesSpecification*, which correspond to the *whole-if-needed* and *part-if-needed* daemons of the frame model (fig. 1) respectively. The domain type of these attributes is an XSL (eXtensible Stylesheet Language) document. XSL integrates a transformation language (XSLT) which enables the definition of rules to transform an XML-document into another XML-document. Thus, it results ideal to specify the inference that will combine or obtain values from the XML metadata of *whole* and *part* metadata records. Besides, this class includes a *constraints* attribute which stores the specification of the constraints (if applicable) that the components of the collection must observe.

And as concerns the instance part of the model in fig. 2, the *KB_Metadata* class represents instances of metadata which conform to a particular *KB_MetadataType*. The specific meta-information of a metadata record is stored in the *specificValues*, whose domain type is an XMLDocument that should conform to the XML-Schema stored in the *syntax* attribute of *KB_MetadataType*. Last, the *KB_AggregationRelation* class is used to describe the instances of the aggregation relations that are established between metadata records. This class includes a *pattern* attribute to identify (if it is applicable) the default spatial/temporal pattern that follow the components. An example where these patterns appear would be the case of geographic information collections that have arisen as a result of the fragmentation of geographic resources into datasets of manageable size and similar scale. Usually, the spatial area covered by the components of these collections follow some type of prefixed division (e.g. the grid establishing the division of tiles at a specific scale or the province boundaries) of the space. Knowing this pattern will facilitate the documentation and organization of the components in a particular collection.

With respect to the dynamic behavior of this model, the most important feature is the ability to infer complete metadata descriptions, ascending or descending through the aggregation relations. The methods presented in *KB_Metadata* and *KB_AggregationRelation* provide the behavior already sketched in fig. 1 for frame facet daemons. Firstly, the method *getCompleteValues* act as the *if-needed* facets but providing the complete values for all the elements making use of the methods *getValuesBeingPart* and *getValuesBeingWhole*. Secondly, the method *getValuesBeingPart* uses, in turn, the methods *getPartDerivedValues* and *getPartInheritedValues* to infer meta-information for a metadata record acting as part.

(Draft) Proceedings of the 15th Database and Experts Systems Applications - DEXA'04 (1st International Workshop on Geographic Information Management - GIM'04): 2004, p. 878-882.

On one hand, the *getPartInheritedValues* method enables parts to inherit meta-information contained in metadata records through the ascending *whole-part* hierarchy. And on the other hand, the *getPartDerivedValues* method enables a part to merge its metadata element values with the values obtained from *getPartInheritedValues* and according to the functions specified in the *partDerivedValuesSpecification* of *KB_AggregationRelationType*. And thirdly, the method *getValuesBeingWhole* makes use of the method *getWholeInferredValues*, which obtains inherent metadata (metadata derived through the analysis of the components of the aggregation) according to the aggregated functions specified in the *wholeInferredValuesSpecification* of *KB_AggregationRelationType*.

Finally, it must be mentioned that *KB_AggregationRelation* also offers special statistics of the elements in the collection: the methods *getItemCount* generates statistics by type or format; and the methods *getSpatialCoverage* and *getTemporalCoverage* generate the spatial and temporal coverage of the components.

4. Conclusions

This paper has presented a solution for the management of nested collection of resources in Digital Libraries, which makes use of XML technologies and knowledge base concepts. It proposes the construction of catalog services over the base of a Metadata Knowledge component. Some of the concepts already existent in ADL and STARTS approaches have contributed to the design of this Metadata Knowledge Base. Similar to ADL and STARTS, the knowledge base makes a distinction between contextual and inherent metadata: the goal of metadata records (*KB_Metadata*) describing collections is to store uniquely specific contextual metadata; and the rest of meta-information, the inherent metadata, is automatically generated by the methods provided in *KB_AggregationRelation*. But our knowledge base component makes two additional contributions to previous approaches. Firstly, it introduces the automatic inference for the records describing the components of the collection, which may inherit meta-information already filled for the collection. And secondly, it gives support for nested collections. In contrast to ADL where it is made a strict separation between collection level metadata and item level metadata, our solution enables the description of collections and components according to the same schema.

Although the concepts presented in this paper are extensible to any type of Digital Library collections, the context of Geographic Information has been used to illustrate the proposals. The management of collections and series of resources is an important need and the knowledge representation model presented here may provide great benefits for the construction of metadata cataloguing systems integrated within Geolibraries or Spatial Data Infrastructures. First of all, this system will avoid the redundancy in metadata creation, metadata is only maintained in one place and inherited whenever is needed. Secondly, it will facilitate the supervision of the metadata creation process by comparing the already catalogued components with respect to the pattern followed by these components. For instance, in the case of a spatial collection, this system will be able to overlap the spatial pattern grid (the division of tiles for a specific scale) and the layer formed by the bounding boxes of the components already catalogued. Another benefit of this system will be the possibility of providing discovery and presentation of metadata records at an aggregated or disaggregated level on user demand. Although not explained here, the knowledge base could deduce whether a initial set of metadata results are describing components of the same collection, i.e. the knowledge base could find the metadata record that subsumes the initial results in the ascending whole-part hierarchy. Finally, the unified description of collections and components can also help to generalize software for access and visualization of aggregated resources.

Acknowledgments

The basic technology of this work has been partially supported by the Spanish Ministry of Science and Technology through the project TIC2003-09365-C02-01 from the National Plan for Scientific Research, Development and Technology Innovation.

References

[1] A. Artale, E. Franconi, N. Guarino, and L. Pazzi. Part-whole relations in object-centered systems: An overview. *Data & Knowledge Engineering*, 20(3):347 – 383, November 1996.

[2] Federal Geographic Data Committee (FGDC). Content Standard for Digital Geospatial Metadata: Extensions for Remote Sensing Metadata. Public review draft, 2002.

[3] K. D. Forbus and J. de Kleer. *Building Problem Solvers*. MIT Press, 1993.

[4] L. Gravano, C.-C. K. Chang, H. García-Molina, and A. Paepcke. STARTS: Stanford Proposal for Internet Meta-Searching. In *Proc. of 1997 ACM SIGMOD Conf.*, 1997.

[5] T. Gruber. A translation approach to portable ontology specifications. Technical Report KSL 92-71, Knowledge Systems Laboratory, Standford University, Stanford, CA, 1992.

[6] L. L. Hill, G. Janée, R. Dolin, J. Frew, and M. Larsgaard. Collection metadata solutions for digital library applications. *Journal of the American society for Information Science*, 50(13):1169–1181, 1999.

[7] International Organization for Standardization (ISO). Geographic information - Metadata. ISO 19115:2003, 2003.

[8] M. Minsky. *Mind design*, chapter A framework for representing knowledge, pages 95–128. MIT Press, 1981.

[9] A. Powell, M. Heaney, and L. Dempsey. RSLP Collection Description. *D-Lib Magazine*, 6(9), September 2000.