In: Computers, Environment and Urban Systems, Pergamon Press, 1993.

AN ARCHITECTURE AND DESIGN APPROACH FOR A GEOGRAPHICAL INFORMATION RETRIEVAL SYSTEM TO SUPPORT RETRIEVAL BY CONTENT AND BROWSING*

M. Agosti, F. Crivellari, G. Deambrosis and G. Gradenigo

Dipartimento di Elettronica e Informatica
Università di Padova
Via Gradenigo, 6/a
35131 Padova, Italy
Phone: +39 49 828 7600
Fax: +39 49 828 7699

ABSTRACT

In present Geographical Systems (GIS) the access to information is implemented for experienced users. On the contrary it is important to permit to final users an easy access to information by content and in a natural way. In fact a final user is not always an expert on geographical information systems, but he is an expert in a specific application area such as urban management or land administration.
Geographical systems need to be able to manage geographical and structured data together with textual data and also data on other media. In fact an active use of textual data can add an effective access point to the collection of geographical information.
So it is necessary to include textual management operations in currently existing geographical systems to provide a useful and comprehensive access point to the geographical collection to non-expert users of the application. This paper introduces an architecture and design approach for a Geographical Information Retrieval System (GIRS) able to support retrieval by content and browsing on textual data; the architecture and design approach gives the framework for the management of various kind of media that is necessary to manage in geographical systems.

1. NECESSITY OF RETRIEVAL BY CONTENT CAPABILITIES IN GEOGRAPHICAL SYSTEMS

Many tools are currently available for management and manipulation of geographical information, but the capabilities of these tools are mainly based upon representations of the complex structure of geographical information. In fact, on these representations numerous operations and operators that permit an effective retrieval and manipulation of geographical data have been defined. From a collection of maps managed by one of these tools it is possible to extract a set of maps that all have an attribute corresponding to a specific value. For example it is possible to formulate a query for retrieval of all land maps that contain land over 200 metres in elevation. This example is truly indicative, because it illustrates the fact that the usual form of retrieval supported by these tools is of deterministic nature. In fact, the operation retrieves all and only those maps that exactly match the condition expressed in the query.

---

* This work derives from the paper, by the same authors, "An Object-Oriented Approach for the Conceptual Modelling of Geographical, Structured, and Textual Data" that has been presented at the 15th European Urban Data Management Symposium (UDMS), 16-20 November 1992, Lyon, France.

In general, geographical applications are implemented by making use of a database management system (DBMS) capable of handling geographical information as well as its related structured data, where the structured data adds significant semantic information to the maps to which it is related. In an application which manages a collection of geographical information the final user is given the possibility to retrieve pertinent geographical information in two different ways, by using:

- operators that are defined and used on the structured data and are usually available in the Data Manipulation Language (DML) of the specific DBMS used, and
- operators defined on geographic information.

Both ways of retrieving information are conducted in a deterministic way and retrieval is done by value. It is usually possible to formulate queries involving information only of a deterministic nature, e.g.: "display all buildings in Venice that have been restored in 1993".

The availability of an access point to geographical information through structured data is rather efficient but it always requires some effort from the user to clearly represent his information needs when formulating a query. The effort of representing his information needs is not so overwhelming if the attributes the user is exploiting to retrieve information identify numerical values or codes, because in this case the possibility of ambiguity would be limited. On the contrary, retrieving information by attributes of textual kind is a very complex procedure as it is for some other multimedia data like images and voice.

Present database management systems do not provide effective and complete management of multimedia data. If we take the textual data as a significative example of the complex data that needs to be managed in multimedia applications, the currently available DBMS operators permit exact match procedures on texts, but they do not permit sophisticated textual operations or retrieval by semantic association.

This means that it is not possible to retrieve multimedia data **by content** or **by association** when the data is managed by a database management system. This becomes extremely limiting when the user is looking for geographical entities basing his research on the semantic content of multimedia data. It is really limiting for all kind of multimedia data, but we are going to concentrate and report in the rest of this paper the results we have reached on the representation and management of textual data, also trough experiments previously developed in different contexts [Agosti et al, 1991a&b, 1992]. The reached results for textual data are going to be extended for other media in future work.

If the user is interested in queries of the following kind: "display all areas that have constrains **pertinent to** environmental protection", any current DBMS would not be able to answer a query formulated in that way; instead a present DBMS is able to answer a query like this one: "display all areas that have constrain equal environmental protection", if and only if the attribute "constrain" assumes the value "environmental protection" in one or more database records. Any other word or term similar in sense with the term given by the user in the formulation of the query would not be suitable for displaying the areas of interest.

The user in the formulation of his query to a DBMS must always use the exact textual values stored in the database in association with a specific attribute. This means that the final user needs to know the schema of the database and he has to use the exact values stored without having the possibility of using synonyms or semantically related terms.

It is necessary to note that a final user is not always an expert on geographical information systems, but he is an expert in a specific application area such as urban management or land

administration. And it is really well-known the existing differences in the technical languages of different application areas: the final user does not always known the terms adopted by the "expert user" that has developed the application or populated the application database.

On the contrary, present information retrieval systems (IRS) are specialised in the management of textual data and in the retrieving of information by content in textual databases.

Geographical systems need to be able to manage geographical and structured data together with textual or other media data. In fact an active use of textual data can add an effective access point to the collection of geographical information.

In many present-day geographical systems the textual data is not automatically managed and related to pertinent geographical data, but in real applications many geographical entities are supported by textual notes, technical reports or legal documents, all textual documents that would be necessary and useful to make usable in an active way by the final user. So it is necessary to include textual management operations in currently existing geographical systems to provide a useful and comprehensive access point to the geographical collection by non-expert users of the application. Here we introduce an architecture and design approach for a geographical information retrieval system (GIRS) able to support retrieval by content and browsing on textual data; the architecture and design approach gives the framework for the management of various kind of media that is necessary to manage in geographical systems (see Figure 1).

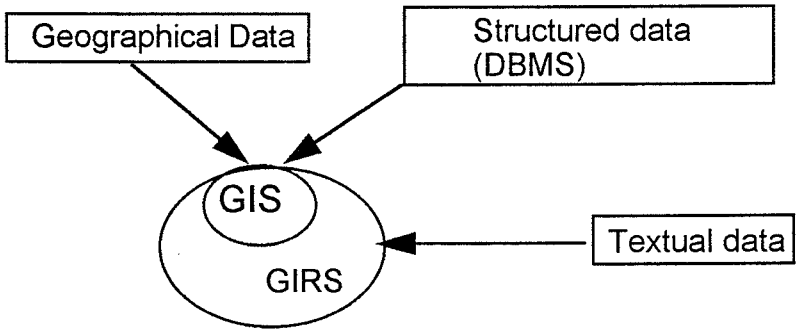

Figure 1. Types of media managed by a GIRS.

2. AN APPROACH FOR THE DESIGN OF GEOGRAPHICAL INFORMATION SYSTEMS WITH INFORMATION RETRIEVAL CAPABILITIES

The design approach for development of geographical information systems with information retrieval capabilities requires the modelling of two different data types: structured and unstructured. In fact data concerning a specific application domain can be modelled at two different abstraction levels, one intensional and the other extensional. These two levels of representation generally reflect the user's conceptualisation of the "reality" to be represented in a data management system. On an intensional level, the formal description of properties and relationships of a set of informative objects, specified in terms of the semantic or logical structure of a particular data base model, is known as the schema. Design and management operations on data can be effectively carried out by using a schema, disregarding the particular implementation. On an extensional level, the objects of the reality of interest described in the schema populate the database.

Data used for information retrieval purposes is a representation of documents. By the term documents we generally mean those objects which contain information of any media and are of

geographical interest: i.e. books, reports, texts of laws, photographs, slides, videos, audio tapes, maps, and so on. It is important to note that at present the information related to and/or contained within these different multimedia objects is usually represented as textual data only. Thus a collection of multimedia documents is represented by a collection of textual fragments incorporating different elements for a deterministic representation of the document; in the following, this representation is referred to as structured data. Other textual data represents the semantic content of each document of the collection; this data is called unstructured data in the following and represents the output of a non deterministic representation process. A system with information retrieval capabilities needs to manage these two types of representation including the connection with the original document that can be of any kind of media. So the data representing each multimedia document is expressed by structured data and unstructured data:

- structured data consists of all those data items which describe the deterministic aspects of the document (e.g. date, place and year of publication);

- unstructured data represents the informative or semantic content of each document; this data is called "unstructured" because it is not possible to manage it on a logical or physical level within a record structure with fields of fixed length. For example, it is impossible to arrange this data with the attributes of a relational database management system.

The existence of these two kinds of data within an integrated database establishes the peculiar character of management of information retrieval data, especially due to the presence of unstructured data. Management of structured data is based upon a consolidated technology: common data processing applications such as accounting or planning are able to manage only structured data. The integrated collection of data generated by applications of such nature consists basically of structured data which can be designed by making use of database design methodologies.

Generally, the kind of data considered peculiar to ordinary information retrieval operations is the so-called **unstructured data**, which is not really "unstructured": it often consists of a collection of terms organised into a semantic structure which is used by the information retrieval system to find the correspondence between terms in the user's query and the documents of the managed collection. The term "unstructured" results mainly from the fact that it is **inadequate** to represent the semantic content of a collection of documents by means of a completely flat set of terms. Different semantic structures have been proposed to manage the collection of terms used by an information retrieval system in order to represent the domain of pertinence of the collection of multimedia documents together with the terms within the structure. The simplest structure used to represent and manage the collection of terms consists basically of an alphabetical list of significant words or terms; other more complex structures can be that of a classification scheme, a semantic network or a thesaurus.

It is possible to introduce a classification of unstructured data due to the nature and application of this sort of data; it is therefore possible to consider two different families of unstructured data: surrogates and indexing terms (also called auxiliary data). Surrogates are basically textual data which often substitute the real document and represent its content within the system; surrogates can be: abstracts, indexes of documents, prefaces, and so on.

Indexing terms (or auxiliary data) is another different technique for representing the informative content of a document. Indexing terms generally consist of a **semantic structure** incorporating a **collection of words** or **terms** which depend upon the domain of the collection of pertinent documents. The indexing term structure identifies the semantic relationships between words or terms: it is this structure which permits implementation of data retrieval operations by content in contrast to the usual technique of data retrieval by value, currently supported by all traditional data processing applications. Many different paradigms and techniques can be used in the design and management of the structure of indexing terms: i.e. object-oriented, knowledge-

based, etc. All approaches have the same basic purpose: to implement a structure which can be used to find semantically related terms; by using related terms which directly concern the user's specific information needs it is possible to launch a query to the database in order to retrieve all relevant documents.

The individual indexing terms represent the vocabulary to be consulted; this means that such a structured collection of data can exist even if the documents have not yet been inserted into the system. For grasping the meaning of the structure to be used to organise the indexing terms and the collection of terms that populate this structure, it can be used an analogy with the schema of a database and the instances that populate the database itself. (See Figure 2)

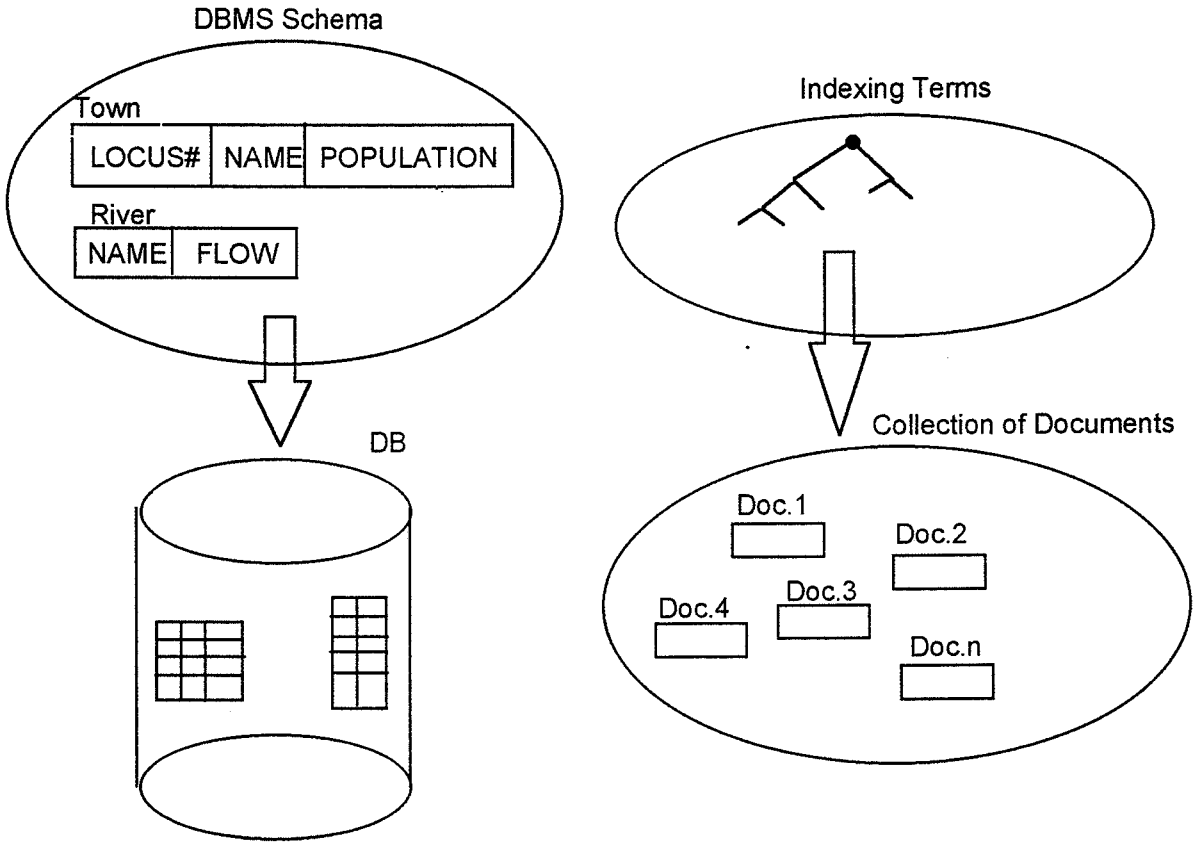

Figure 2. Analogy between Schema/DB and Indexing Terms/Documents.

The indexing terms that populate the semantic structure are associated to each document by means of an indexing process in order to represent the semantic content of each document in the database. Each indexing term is assumed to describe the document content only to a certain extent, neither completely nor uniquely. Various different sets of indexing terms may be assigned to the same document; in fact, the description of the content of a specific document does not arise from a deterministic process as in the case of surrogate data. Furthermore, the content can either be described in different ways by several different people or the different descriptions simply reflect different users' requirements for information.

Various different models have been proposed for the management of indexing terms [Belkin & Croft, 1987; Salton & McGill, 1983; van Rijsbergen, 1979]. These models differ considerably from one another, mainly due to the different structure of the indexing terms.

Some models permit only an exact match between the user's query and the unstructured

data, these models are capable of managing surrogates only. These models make use of a representation of the document informative content actually given by the surrogates or, more simply, a list of words associated to the documents. It is important to note that a list of alphabetically ordered terms contains no semantic relationship between its component words, thereby it merely represents a flat and semantically poor structure.

The models based on a partial match retrieval technique are also capable of operating an exact match, so these models have all the characteristics of exact match models, furthermore they normally use a specific indexing term structure. Depending on the indexing terms structure identifying the model, different types of systems can be implemented, one of these models being the probabilistic one [van Rijsbergen, 1979]. Some other models based on hypertext techniques are currently being developed and have been presented in [Agosti et al, 1989, 1991a&b, 1992; Croft & Turtle, 1989; Lucarella, 1990].

3. REASONS FOR AN OBJECT ORIENTED APPROACH

To reach the target of achieving such capabilities from a Geographical Information Retrieval System (GIRS), several different approaches can be adopted. In particular, one possibility would be to expand the relational data model, at present the most commonly used traditional DBMS. But this model imposes the first normal form limitation. This means that the object space must be mapped onto a collection of flat relations and this limits the expressive capacity of the model. The access to data is based only upon values and the concept of entity identifier is not used. This reduces the possibility of introducing relationships between entities due to the fact that the relationships are not supported directly by the model. Furthermore an extended relational model would not be capable of using non-traditional domains (i.e. an attribute defined on a textual domain) in a completely coherent way with traditional domains. For example, such a model would not be able to express the relationships existing between the terms that represent the informative contents of geographical entities.

These relationships between terms form a semantic structure to be used in describing the informative contents of geographical entities and the same structure should provide the user with a frame of reference in the query formulation process. The relational model does not directly support the abstraction mechanisms that are necessary in modelling geographical entities and the structure of terms to be used in describing the contents of these entities. The abstraction mechanisms necessary to organise structured, spatial and textual data in a geographical application, are the classification, specialisation- generalisation and aggregation mechanisms.

The classification mechanism is a very fundamental and intuitive one. The mechanism permits the grouping of the documents that share one or more properties of structural or semantic nature. The application of the classification mechanism to geographical entities produces classes of objects of the same type (e.g. maps and textual documents) and classes of objects (usually of different types) that are related to the same subject. A class and its instances are related by means of an "instance-of" relationship; the instance-of relationship between a document and a class is implemented by associating the term which identifies the particular class to the document. In a GIRS, the class definitions provide the end user with a description of the database's informative content. The classes delimit the specific area in which the system is recognisable and define the vocabulary by which the user accesses the information. For classification of geographical information retrieval (GIR) data, it is essential to use a polythetical classification scheme [van Rijsbergen, 1979].

The generalisation-specialisation mechanism [Smith & Smith, 1977a] simply relates a set to its subsets or a class to its subclasses. The representation of the generalisation- specialisation mechanism is usually expressed by a "subset-of" or "subclass-of" relationship. An example can be given by examining the relationship existing between a set of documents on the topic of

environment morphology and the subsets of that set of documents that can be created by distinguishing the documents on the different aspects of environment morphology. The different aspects could be: hydrography, orography. etc. These relationships can be seen as a hierarchical structure that organises sets and subsets or classes and subclasses.

The aggregation mechanism [Smith & Smith, 1977b] relates objects of the same or of differing types by transforming a relationship between several objects into a single object of a higher level. This new object can have specific, individual characteristics of its own. The aggregation mechanism is used to model the user's perceived relationship between concepts. For example, this abstraction allows definition of the relationship between the terms "environment" and "morphology" used separately, taking the term "environment morphology" as a distinct concept.

An approach which could overcome the limitations of the relational model in the representation of the semantics of textual data and the direct expression of the abstraction mechanisms previously introduced is the object-oriented approach. The basic concept of the object oriented approach is the idea of an "**object**". An object is anything that may have its own identity or singularity, whether it may occur in the physical world or in the world of concepts. Thereby an object can be either a physical object or a conceptual object accordingly. The object notion is strictly connected to the notion of the object's "**identity**": an object can exist independent of the fact that we might regard it as having its own characteristics or properties. It is possible to establish relationships directly between objects without any reference whatsoever to the object's properties. As a consequence every object can be given individual identity without reference to the values of its properties; the object's properties may change with time whereas the identity of the object persists. These object-oriented features provide greater modelling flexibility and power than the traditional record-type-oriented model. In a GIR context such object-oriented features allow an object to be inserted into the database even when its properties or its relationships with the indexing terms are not known. This means that it may be possible to construct the indexing term structure without having inserted any data from the collection of documents; a system set up in this manner might be able to actively and independently manage the indexing terms and the collection of documents.

The object-oriented approach, used for modelling purposes, is able to support direct representation of two essential data abstractions, namely classification and generalisation-specialisation, providing a natural framework for GIR data modelling. The classification abstraction mechanism groups together similar objects into a **class**; the generalisation- specialisation mechanism can be applied in order to model the relationships between classes producing a taxonomy of classes. The classes of a taxonomy are related by superclass/subclass relationships. A subclass inherits all the properties of the superclass and other properties can be added, if necessary, in order to specialise the subclass.

A class may be considered itself as an object, from this point of view some properties are intrinsic characteristics of the class itself as an object (i.e. a property of the class) and other properties are common to all objects which are elements of the class (i.e. a property of the objects). By perceiving classes in their turn as objects, the classification mechanism can be applied to classes so that the classes can be grouped together into "**meta-classes**".

For the organisation and the associative access to a real document collection it is rather commonplace to use a classification scheme as a hierarchical structure. By means of a classification scheme each document of the collection is classified into a subclass which groups together all documents pertinent to a specific aspect. Thus a classification scheme can be seen as a mechanism by which a hierarchical tree structure of subclasses of documents can be built. The collection of such documents is therefore transformed into a well-organised, tree structured order.

4. ON THE REPRESENTATION OF THE SEMANTIC CONTENT OF MULTIMEDIA DATA

The relational model and the Entity Relationship approach which are widely used for the design of structured databases do not permit the representation of the complex objects that need to be represented in a GIR system. The typical object that needs to be managed by a GIR system is a complex object that contains data of many different media, and the GIR system has to support connections between the different parts of different media of the same object. It is important to note that when a GIR system manages these complex objects, each object has to be retrieved using each of its component media as a semantic entry point.

Another important function that needs to be supported by a GIR system is the capability of retrieving complex objects using a semantic access on the different media. To retrieve the complex objects by a semantic access means that each of the media of the object needs to be indexed. It is important to note that the indexing process corresponds to the application of the classification mechanism to documents.

At present, algorithms and procedures of automatic indexing of textual data are of common usage in the information retrieval area and can be imported in a GIR system, whereas the automatic indexing of different media is an open problem, because only initial results exist for the indexing of geographical data and pictures. In particular, initial results have been reached for the content on image representation in image databases [Rabitti & Savino, 1991]; the approach by Rabitti & Savino has the target of representing complex objects which are present in the images, that is the use of the classification mechanism previously presented in this paper to the images. It is important to note that this automatic classification process at present can be applied to images only if the domain of the application is determined and described in advance.

A consequence of this situation is that the indexing of collection of geographical data can be only manually made by professionals in the application domain and in the available GIR system.

The indexing process is important because it permits the identification of relevant indexing terms of the documents to be used in the semantic organisation of the document collection. The indexing terms are not semantically equivalent, but among them there exist different relationships, because some indexing terms are specific, other general, and in other case some are equivalent. This means that a collection of indexing terms needs to be organised in a semantic structure expressing the existing semantic relationships. It is important to note that a generic relationship between two indexing terms corresponds to a set relationship between the two sets of documents respectively related to the two indexing terms. (See Figure 3)

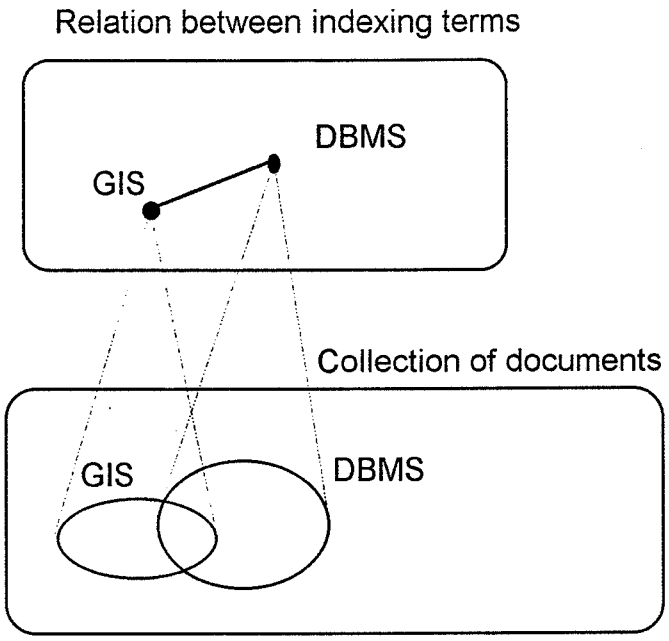

Figure 3. An example of the relationship between indexing terms and sets of documents.

5. THE CONCEPTUAL ARCHITECTURE

The use of an object-oriented approach permits the definition of an architecture for the conceptual representation and design of GIR data, whereas the actual GIR system needs to be able to manage complex objects such as a document of a collection of multimedia documents or as a structure and a collection of indexing terms that semantically represent the collection of documents.

In this approach each indexing term is viewed as a class of objects; instances of each of these classes are the documents which are pertinent to the specific term expressed by the class; this "term" represents the class and corresponds to the concept identified the class. Hence one class is seen as a set of documents and from a higher abstraction level may be seen as a single conceptual object of a structure representing the semantic relationships between different concepts.

Each class can be perceived from two different abstraction levels:

- as a set of instances,
- and as a whole entity.

The class as an entity can incorporate properties of its own; these properties are assigned to the class as an object, that is the properties of the concepts as such. It is important to note that these properties of the class must be separated from the properties of the individual objects that are instances of the class. In our design approach a class denotes a concept, thus the concept has its own properties that are distinguishable from those of the individual documents semantically represented by the concept. One must bear in mind that this approach implies working on at least three different classification abstraction levels.

First level

This is the lowest level which contains the objects of interest: maps, texts and tables. It is

the level where the collection of documents of a GIR application is stored. Each document of the collection of documents has its own identity and status; the identity of the document is independent of the manner in which it is represented or structured and of the values it may assume. It is important to point out that this possibility is in direct contrast to the widely used relational approach where a tuple can exist only if the relationship (i.e. the structure capable of receiving the tuple) has been previously defined and if the values of the attributes have been assigned to the tuple.

Second level

This level arises from the application of the classification abstraction scheme to the objects of the first level; this level can be identified as the level of concepts, that is, the locus on which the semantically related concepts are placed; this is the plane of abstraction where the terms used in a GIR application can be listed; each term identifies a concept that is pertinent to a certain set of documents on the basis of the semantic content and these documents belong to the class identified by the specific term. The conceptual tool necessary for designing GIR data must be able to support polythetical classification. In this kind of classification every component of one class receives only a portion of all the attributes possessed by members of the same class; hence no attribute is both necessary and sufficient to determine the membership of any one element to a class. For example, a map is usually pertinent to some specific territorial aspect and therefore needs to be indexed by several terms, each term identifying a class to which it should belong. This conceptual tool is usually identified by a structure and a collection of indexing terms.

Third level

The third level of the classification abstraction hierarchy is the plane of meta-classes; a meta-class is a class of classes, in fact its members are classes of elementary objects; a meta-class can be, for example, a classification system or vocabulary used in a GIR application to index a set of documents.

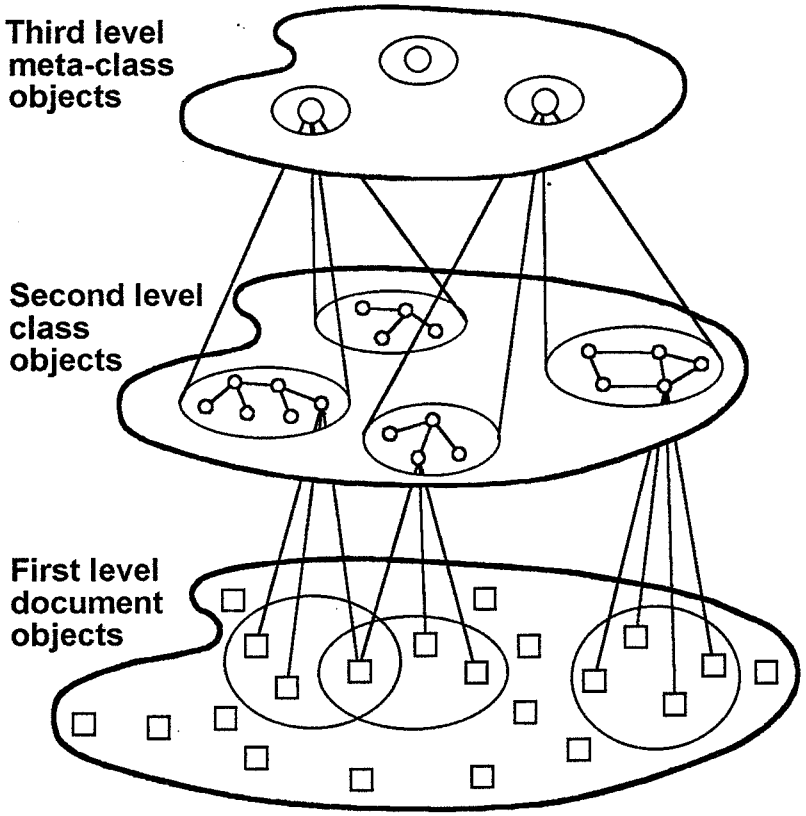

Figure 4. The three level architecture representation.

In the following we present the aspects of our design approach in detail; the third level is not pertinent to this work, furthermore it is not here addressed.

5.1 First Level

The collection of interest is represented and managed at the first level of the architecture. The collection is made of objects, an object is made up of a set of fragments of the same or of different types and fragments are connected by structural links.

The representation of an object of this level can be accomplished by using:

- multimedia fragments of a document, such as images (raster or vectorial data), alphanumeric structured data, or textual data;
- documents that make up the conceptual units for indexing, accessing and retrieving.

A fragment of a document can be shared by different documents; for example, a specific map of a city can be a fragment of these different documents, the project of a new building of that city or the project for the new planning of the road network of that city.

The documents can be mutually connected so as to express the semantic relationships between them, each connection is made by using reference links that establish a network of cross-reference links between documents. In this way the documents are interconnected within a network at the first level of the architecture, then the collection of documents is arranged on the first level of the architecture as a network of structural links combined with the network of reference links; this complex network is called "hyperdocument".

Structural links are used to combine the different fragments to make a whole document. The structural links model the hierarchical structure existing between the multimedia parts of a real document. An example of the structure of a document is represented in Figure 5.

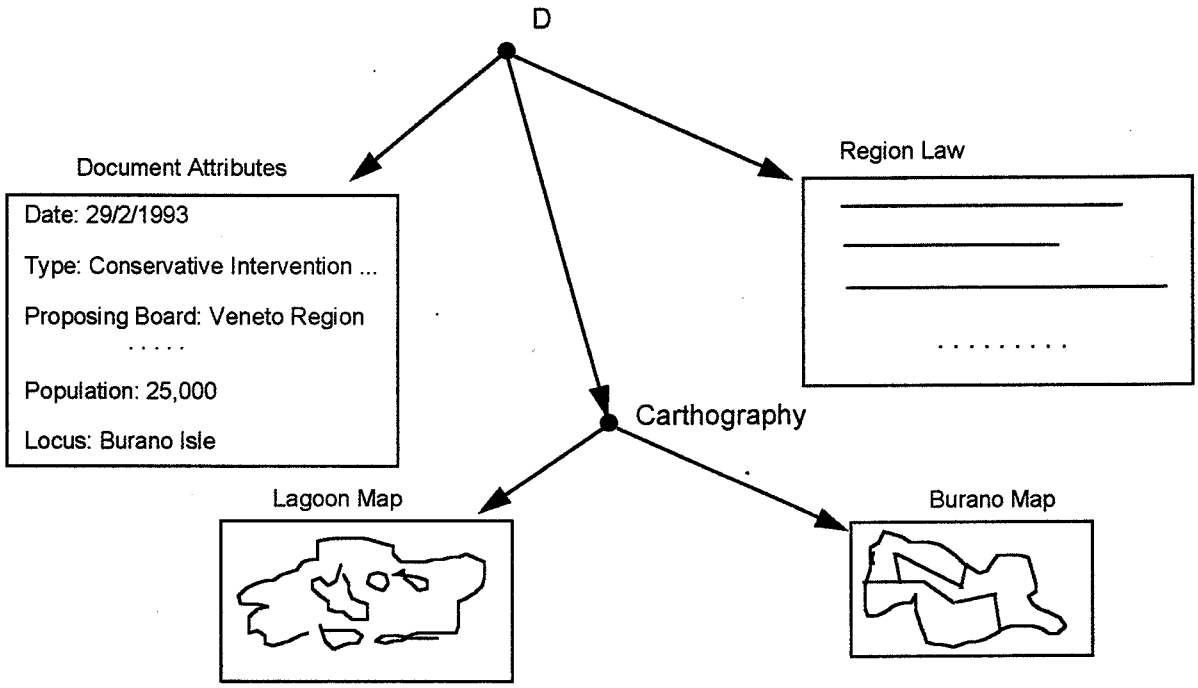

Figure 5. Structure of a hypothetical document concerning the project for a conservative intervention on the isle of Burano (Venice lagoon) approved by the Veneto Region authorities on 29 February 1993.

The structural links are obtained by connecting a father node to its offspring in order to form a branching diagram within the global multidimensional diagram of the hyperdocument. It is important to note that these links are set up by the designer and administrator of the hyperdocument; the final user can merely use them.

Reference links allow the semantic relationships existing between different fragments of the same or different documents to be made explicit and represented. Reference links can also be handled by the final user according to his information requirements in relation to linking of documents. In this structure the direction of a reference link has particular meaning; the significance of such possibility is evident if we consider the difference in the semantics for a text being referred to, or for one referring to another. An example can be a fragment of a law that makes reference to a specific article of another law (see Figure 6).

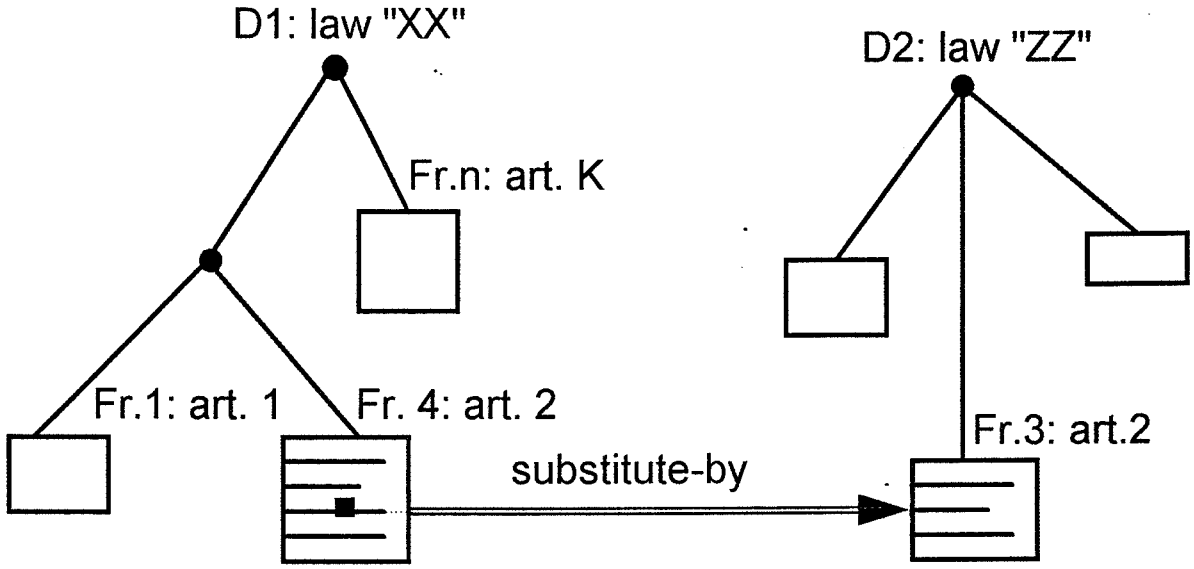

Figure 6. Example of a reference link between articles of two Italian laws.

At the first level of the architecture the approach supports navigability through the document collection, this type of navigability is characteristic of hypertext systems.

Since specific cross-references often exist between the documents of the collection, the system needs to be able to support navigability through these connections. Furthermore, assessment of one item of information generally stimulates request for other information in further depth. The implementation of a hypertext network between the various information items permits their direct consultation. To reduce the common problems of disorientation and knowledge overload which face the user during the use of the hyperdocument, a simple searching technique for detection of text strings located within the full text information items has been introduced. In fact the opportunity of pin-pointing with a certain approximation the whereabouts of some nodes and to use them as starting points for one's own queries has been considered appropriate.

5.2 Second Level

The indexing terms are represented and managed at this level of the architecture. In order to effectively design and manage indexing terms, the architecture offers constructs supporting the fundamental abstraction mechanisms previously introduced together with some other semantic relationships that are necessary to explain specific relationships between concepts, such as, for example, the synonymy between concepts. In fact, the synonymy can be represented by a relationship made explicit between two concepts.

Everything existing in the application world, even those linguistic entities used to describe other entities, are modelled as objects. All abstractions, having as their purpose the conceptual organisation of objects, are represented as links between objects. Therefore at this level different types of links are present. Objects of the second level result from the application of the classification abstraction mechanism to the objects of the first level; they denote concepts which are interrelated, for example through a classification hierarchy, a specific case of which may be the IS_A hierarchy. For example the indexing term "pollution" can be used to represent the set of objects of the first level (documents) regarding chemical and noise pollution. In this case the indexing term is used to represent a qualitative characteristic of the objects. But an indexing term can also be used to represent physical characteristics of the environment, in fact the indexing term "hill" can be used to

represent different objects such as maps, that contain hills and reports describing a specific group of hills.

Objects and links of this level form the "hyperconcept", that is, a parallel structure to the "hyperdocument", whose task is to handle the semantic structure of concepts (indexing terms) used to describe the contents of the document collection. This level is conceptually located above the hyperdocument and performs the same functions performed by the usual indexing terms of an operative information retrieval system.

The process of associating an indexing term to a document object, in order to describe its informative content, corresponds, in this architecture, to setting up an instance-of relationship between a class term (i.e. an object of the second level) and a document, which represents an object of the first level. Being a document generally indexed by more than just one specific term, a document object proves to be an instance of various different term classes. These modelling capabilities support the notion of polythetical classification.

The representation of the informative content of first level objects is to be done for each complete object. This means that each multimedia document has to be considered as an atomic object. Therefore the representation that is inserted at the second level is for each unbroken document. The single fragments are not specifically indexed. Each whole document is linked to its semantic representation in the second level with a "pertinent-to" type link.

One can imagine each indexing term as defining a set of first level objects: the documents which are a specific "instance-of" indexing term, that is the documents indexed by that particular indexing term. On the other hand, document sets defined by two or more distinct terms are not necessarily disjointed so the common elements of the intersection are documents which belong to different classes. While links between concepts and document objects model the classification abstraction mechanisms, links connecting conceptual objects can express generalisation and specialisation abstraction mechanisms. If two concepts are related by a specialisation relationship, the two corresponding sets of documents at the first level are associated by a sub-set relationship; the relationship between sets of documents is an inclusive-set one if the concepts are related by a specialisation mechanism. As for the generalisation-specialisation relationships, all the structuring mechanisms of indexing terms correspond on the first level to a set structure of the collection of documents (see Figure 7).

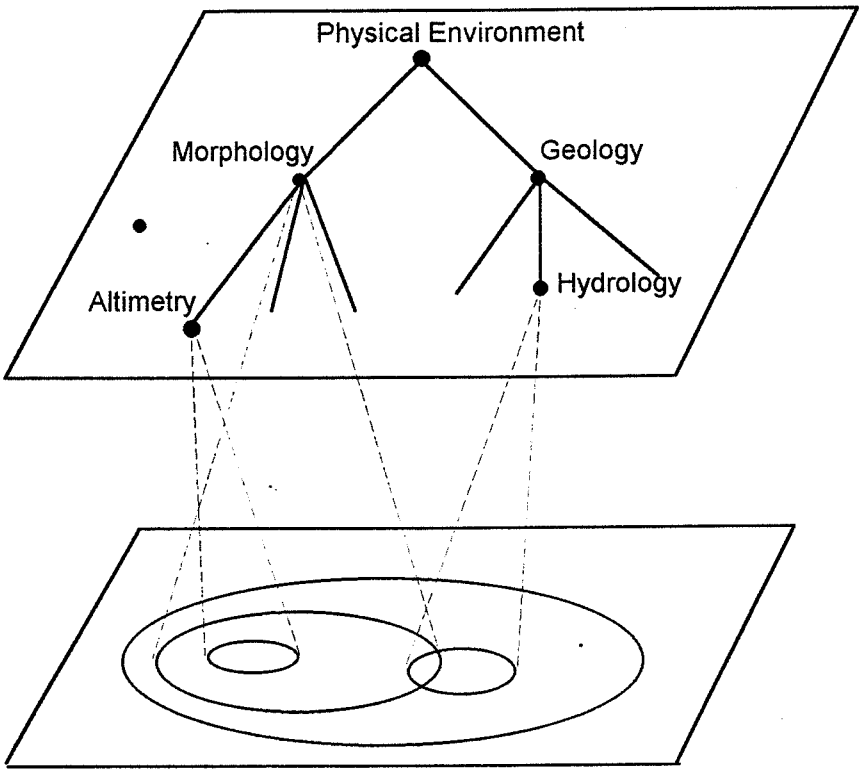

Figure 7. Impact on the documents of the generalisation/specialisation mechanism.

On the second level, navigability through the semantic structure permits formulation of a query by identifying a semantic path through the reference structure. The use of simpler structures (i.e. classification schemes) or of more complex nature (i.e. thesauri) has no essential significance in determining the construction of the mechanism. An example of navigability through the semantic structure is shown in Figure 3. We can imagine the user is navigating through the semantic structure and visits the nodes representing these concepts: "physical environment", "morphology", and "altimetry". If he expresses his interest in seeing the documents related to the path represented by those concepts, a system based on this model has to be capable of identifying the set of pertinent multimedia documents. This can be possible by making use of links created by applying abstraction mechanisms between concepts and documents. The extracted set of documents can be seen in different ways by the final user: it can be accessed directly, or the system can present the set after having ordered the documents by descending values of pertinence with the terms used in the path by the user. When the user has accessed a specific document, he can continue to browse the set of selected documents, or he can start to browse the hyperdocument using the reference links. In this second mode of perceiving the documents, the user can leave the set of documents selected by the system and he can start a different browsing over the hyperdocument.

The independent nature of the two levels of the architecture allows us to take a step further, that is, it provides us the opportunity to construct different and distinct hyperconcepts upon the same hyperdocument. In this way it is possible to obtain different semantic descriptions of the same document collection, that is different views for different categories of users. This feature is quite significant, because a user specialised in a specific field tends to use different terminology compared to that used by a generic user. This means that the approach allows us to construct different access mechanisms and different types of user interaction according to the different access requirements of the various categories of users.

5.3 Relationships between First and Second Level

This section describes the relationships necessary between the first and the second level of the architecture and the operations which need to be supported by the approach. Each of the two levels of the architecture represents a distinct network of nodes and links. The relationship between the terms included in the hyperconcept and the related documents present within the hyperdocument are described by a peculiar type of link. This type of link makes it possible to set up a mechanism of access to the information items, for example, by means of a thesaurus of the hyperdocument domain.

According to this approach, the hyperconcept and the hyperdocument which make up the two levels of the architecture are independent from one another. This means, for example, that the insertion of a new indexing term into the hyperconcept does not imply any modification of the hyperdocument; in the same way, insertion of a new document doesn't entail any variation in the hyperconcept; the only consequence is an activation of new connections between the hyperconcept and the hyperdocument.

The model supports navigation between the two levels by means of the navigability function [Agosti & Marchetti, 1992]. In this way it is at all times possible to pass from the hyperdocument to the hyperconcept and back again. Such freedom of movement requires the support of an appropriate function to make user interaction easier. The approach also has to support a backtracking capability. With this capability the user is supported in finding his way back, step by step, along the path from whichever point in the network he has reached. This is important in that it lessens the requirement for user know-how, imposed by the presence of various alternative paths during navigation. This function is also supported for the reason that the user has to be able to take any path confidently, without being afraid of losing reference, for instance with other alternative directions. Backtracking should be possible for any path within as well as between the two levels of the architecture.

6. FUTURE WORK AND CONCLUSIONS

At present some experience has been gained from development of two different prototypes implementing aspects of the approach. The approach has proved to be effective in the management of textual and structured data. Future work is addressing the development of a prototype in an object-oriented environment to prove the complete approach as a basis of a Geographical Information Retrieval System.

ACKNOWLEDGEMENTS

This work has been supported by the Italian National Research Council (CNR) under the project "Sistemi informatici e calcolo parallelo - P5: Linea di Ricerca Coordinata MULTIDATA".

REFERENCES

[Agosti & Marchetti, 1992] M. Agosti, P.G. Marchetti. User navigation in the IRS conceptual structure through a semantic association function. The Computer Journal, 1992, 35 (3), 194-199.

[Agosti et al, 1989] M. Agosti, G. Gradenigo, P. Mattiello. The hypertext as an effective information retrieval tool for the final user. In: A.A. Martino (Ed), Pre-proceedings of the 3rd Int. Conf. on Logics, Informatics and Law, Vol. I, Firenze, 1989, 1-19.

[Agosti et al, 1991a] M. Agosti, G. Gradenigo, P.G. Marchetti. Architecture and functions for a conceptual interface to very large online bibliographic collections. Proc. of RIAO '91 Conf. "Intelligent Text and Image Handling", Barcelona, Spain, 2-5 April 1991, Vol. 1, 2-24.

[Agosti et al, 1991b] M. Agosti, R. Colotti, G. Gradenigo. A two- level hypertext retrieval model for legal data . In: A. Bookstein et al (Eds), Proc. 14th ACM-SIGIR Int. Conf. on Research and Development Information Retrieval, Chicago, USA, 1991, 316- 325.

[Agosti et al, 1992] M. Agosti, G. Gradenigo, P.G. Marchetti. A Hypertext Environment for Interacting with Large Textual Databases. Information Processing and Management, 1992, 28 (3), 371-387.

[Belkin & Croft, 1987] N.J Belkin, W.B. Croft. Retrieval techniques. Annual Review of Information Science and Technology (ARIST), 1987, Vol.22, 109-145.

[Croft & Turtle, 1989] W.B. Croft, H. Turtle. A retrieval model incorporating hypertext links. Hypertext '89 Proc., Pittsburgh, Pennsylvania, 1989, 213-224.

[Lucarella, 1990] D. Lucarella. A model for hypertext-based information retrieval. N. Streitz, A. Rizk and J. Andre' (Eds), Hypertext: concepts, systems and applications, Cambridge University Press, Cambridge, 1990, 81-94.

[Rabitti & Savino, 1991] F. Rabitti, P. Savino. Image query processing based on multi-level signatures. In: A. Bookstein et al (Eds), Proc. 14th ACM-SIGIR Int. Conf. on Research and Development Information Retrieval, Chicago, USA, 1991, 305-314.

[Salton & McGill, 1983] G. Salton, M.J. McGill. Introduction to modern information retrieval. McGraw-Hill, 1983.

[Smith & Smith, 1977a] J.M. Smith, D.C.P. Smith. Database Abstractions: Aggregation and Generalisation. ACM Trans. on Database Systems, 1977, 2 (2), 105-133.

[Smith & Smith, 1977b] J.M. Smith, D.C.P. Smith. Database Abstractions: Aggregation. Comm. of the ACM, 1977, 20 (6), 405-413.

[van Rijsbergen ,1979] C.J. van Rijsbergen. Information Retrieval (2nd Ed). Butterworths, London, 1979.