ASSISTING USERS IN SELECTING AND RESTRUCTURING DATA SETS

Sandrine BALLEY
Laboratoire COGIT - IGN
2, avenue Pasteur, 94 165 St Mandé Cedex
France
sandrine.balley@ign.fr

ABSTRACT

This paper addresses the issue of geographic data fitness for use. We want to assist users in finding data precisely adapted to their application's requirements. An approach is proposed to provide customised data sets, thanks to a system helping users to choose and to restructure existing data. This system mainly relies on a data set description model. This paper describes the issue and chosen approach for this beginning research project, but no result is presented yet. The core elements of the data set description model are presented at the end of the paper.

KEY WORDS

Geographic Data Access, Fitness for Use, Data Set Customisation, Data Set Description Model.

1. INTRODUCTION: USER ACCESS TO FITTED-FOR-USE GEOGRAPHIC DATA SETS

1.1. GENERAL IDEAS

Users of geographic data are not always experts in the geographic information domain: they may not be aware of the multiple slight differences making each geographic product more or less adapted to their application.

Some research work has been carried out to customise geographic processing or GIS tools [1] [2]. On the other hand, with regard to data sets, users up to now have had to do with rigid products which are proposed exactly as they were designed. However, user requirements are various, even concerning one and the same data set, as can be noticed in a pricing experiment described in [3].

This paper introduces a starting research project that aims at proposing data sets customised for user need, thanks to an interactive specification of data extraction and restructuring.

Important research has been carried out concerning user access to data set description, especially through metadata [4]. The SDI Cookbook [5] defines three levels of metadata for spatial data infrastructures:

- The discovery level provides global information introducing the geographic product. It should enable the user to know which products exist.
- The exploration level gives details about data sets composing the product, and should enable to know whether the data will meet general requirements of a given problem.
- The exploitation level provides information required to load and use the data in the final application. It includes a data dictionary, the data schema, reference system and geometric characteristics, etc.

As shown in Figure 1, the metadata level we are focusing on is located on the exploration level and on part of the exploitation level. The user we consider is aware of existing data products (the discovery step is over). He wants to select among one of those the data containing the information specifically required by his application. He also needs to adapt the selected data, so that the information has the adequate modelling and can be used for the planned application as soon as the data are distributed.

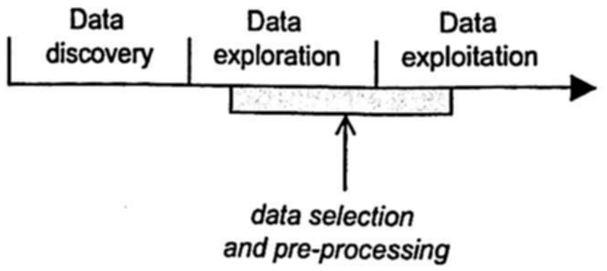

Figure 1. Levels of metadata in user access

The following example illustrates the process of user access and points out encountered difficulties.
A risk manager is looking for geographic data supporting flood simulations. For this purpose, he needs to have a representation of bridges, rivers and their crossing points. He has heard about a vector database called GeoBase and

wants to check if it suits his requirements. Several questions must be answered.

- Does the information content correspond to the required content? To answer this question he checks the list of GeoBase feature types found on a metadata server: GeoBase does not provide any "bridge" feature type, nor any attribute domain including "bridge" or any synonym. However, had he been able to read the product specifications, he would realise that bridges are represented in GeoBase: a *Road-Section* feature whose attribute *level* has value 2 corresponds to a road bridging another feature. To know if it bridges a river or something else, the geometric intersection must be tested with *River-Section* features. So, the information is represented in an implicit way. It can be made explicit through a simple operation available in most GIS software. Moreover, the user cannot find any "road" feature type. Information is geometrically divided into "road sections", and represented by a large number of feature types distinguishing main roads, motorways, cycle tracks, etc.: GeoBase contains the required information, but its level of detail exceeds the user need.
- Is the data schema adapted to the application? In our example, it is not. However, the schema can be modified by some schema transformation operations: the *Road-Section* feature class can be split up by filtering the values for *level* attribute, so that the *Bridge* feature class appears. The *River-Section* features can be aggregated to generate *River* features. Moreover, the risk manager has some classical applications to perform: the *River* feature selection and join (to add some water level data), and the path processing on rivers. As GeoBase is a vector database, feature selection and join are enabled. On the other hand, rivers in GeoBase are surface features and do not constitute a network. The application is not feasible on GeoBase in its present state.

This example shows that describing a data set content and fitness for use is far more complex than listing feature types. The next two subsections further describe the problems of data selection and pre-processing.

1.2. ISSUES INVOLVED IN SELECTING ADEQUATE DATA SETS

The selection of a data set implies first of all the selection of an information content: a user specialised in water applications is likely to select the data set providing the most detailed representation of water bodies, whatever structure the data set has. The possibility to extract only needed information from a data set contributes to its fitness for use.

A first difficulty hampering selection is the lack of detailed and available descriptions of data sets content. Sections 1.2.1 and 1.2.2 expose this problem by distinguishing content descriptions provided by metadata and by data sets technical documentation.

1.2.1. Using metadata to select adequate data sets

Metadata are the main way for users to explore available data sets, but the description provided by metadata standards do not particularly focus on the product content [6]. Metadata above all describe the data set in its globality (e.g. its global quality and spatial extension), and do not insist on the individual description of data set elements. The ISO 19115 standard [7] already provides three interesting metadata entities in this context: a subset of feature types occurring in the data catalog (MD_FeatureCatalogDescription), an application schema in a graphic file (MD_ApplicationSchemaInformation) and an image file illustrating a sample of the data set (MD_BrowseGraphic). However, all of them are not mandatory: most of the time, data providers only fill the features type list. Moreover, these metadata entities provide limited information: geometric and semantic representations chosen for the data set features are not detailed.

1.2.2. Using description provided by technical documentation to select adequate data sets

The complete technical documentation written by the data producer (data schemas and product specifications) could provide potential users with more information, but, as we explain in this section, this technical documentation is not distributed and anyway would be difficult to interpret.

1.2.2.1. The technical documentation form

Data schemas.

Database terminology distinguishes three types of schemas for data description, from the most abstract to the most concrete point of view [8].

The conceptual data schema defines information represented by the data and the logical relationships organising this information. It can be used to communicate database content to users, independently of the way it is modelled or stored in a computer: it is platform-independent. In this paper, the feature types, relationships and their attributes defined in the conceptual data schema will be globally called "representation elements".

The logical data schema describes the data structure following the model of a particular database management system (DBMS). It lists tables and key attributes implementing the representation elements of the conceptual data schema.

The physical data schema describes the system of files adopted for data storage on the computer.

Figure 2 represents the three types of data schemas as the successive steps of an abstraction process translating real world features into computer representation.

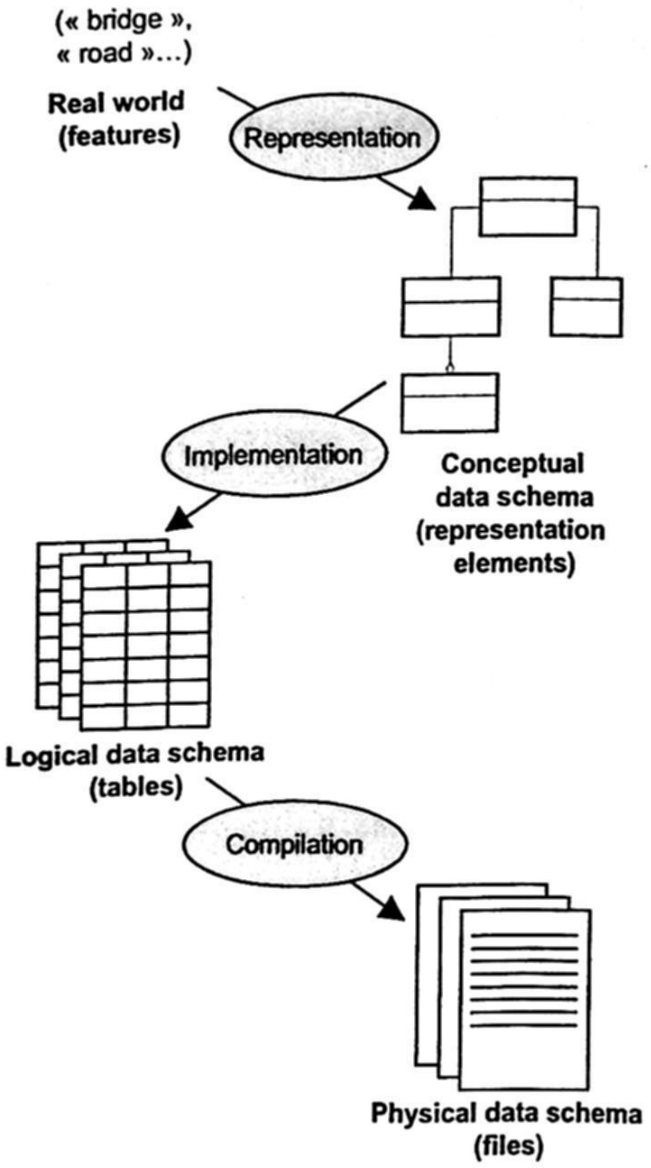

Figure 2. Real world and levels of data modelling

The conceptual data schema is the most relevant to assess the information content of data. However, it does not precise exactly what real world features are represented by representation elements, and how (e.g. what real world buildings can be included in the representation element called "building"). Indeed, real world features have different meanings for every user and representation elements have different definitions in every conceptual data schema. No universal ontology is able to link these two levels [9]. More precision about what information content is provided by a class of the conceptual data schema can be found in the specifications of the data set.

Data specifications.

The specifications provide conditions on real world feature characteristics to select those that should be represented by each representation element of the data set (e.g. buildings whose ground surface is more than 5m² and lower than 15m² are captured as "huts"). They also explain how the features must be represented (e.g. "huts" are represented as points indicating their centre). These textual conditions and rules are numerous and hardly readable by an end-user. Several research approaches are trying to make this documentation accessible to humans or machines. Gesbert [9] and Rüther [11] formalise specifications, which are usually complex text documents, into conceptual data schemas. Their goal is to make data sets unification possible. Goder [12] has proposed a cartographic representation of data sets specifications, insisting on differences between products. As shown in Figure 3, tables (at the top) compare what concepts are represented and distinguished in the data sets. Representation rules are graphically displayed (at the bottom, for rivers of different width).

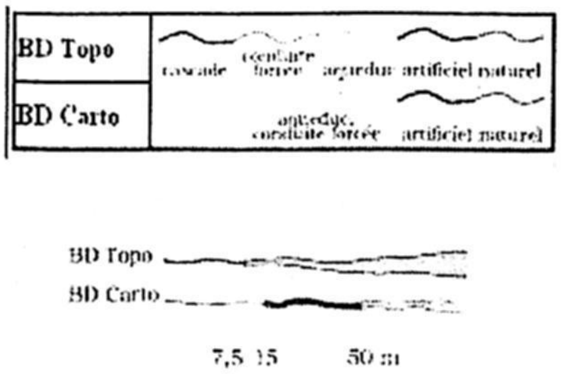

Figure 3. A user-friendly display of content and representation specifications for two IGN databases (BD Topo and BD Carto).

1.2.2.2. Conceptual data schema complexity

A second difficulty hinders data selection in addition to the difficulty involved in reading the technical documentation: it is the intrinsic complexity of the conceptual data schema. The conceptual data schema does not only represent the data set information content independently of other concerns: some representation and implementation choices are also implicated. The main choice is to model information in terms of homogeneous entities, even if it entails decomposing real world features. This is done in order to simplify data capture, storage, and management for the data producer.

For example, in some IGN data conceptual schemas, roads are composed of road sections, which are the longest sections without any intersection and with homogeneous attributes value. This choice is very useful: it renders the road network explicit, and it enables changes in attribute values along the road. However it does not necessarily make sense for users just checking if roads are represented in the data base.

To give a second example, in some IGN conceptual data schemas, distinct representation elements (such as "linear construction" and "surface construction") may represent the same real world feature (such as the concept of river embankment). It is a modelling choice made with data storage in mind, since point, line and area objects must be separated.

These modelling choices suit very well to data management, but they are not the most adapted to the data set content discovery and selection, which most of the time requires a lower level of detail. In the conceptual data schema, the representation elements are very numerous, and their names and hierarchical organisation do not directly reflect the basic real world features required by the user.

1.2.2.3. Conceptual data schema insufficiencies

A third problem makes conceptual data schema insufficient for data selection: not only does the conceptual data schema express the data set information content in a complex form, but also it does not express the whole information content. This is due to the importance of implicit information. As shown in the example of section 1.1 concerning the "bridge" feature, the conceptual data schema (and specifications) does not tell everything: it only describes features that are explicitly stored in tables. In fact, in some cases, simple needed concepts are provided somehow by the data set, but they are not explicitly specified in the conceptual data schema:

- In some cases, the concept has been aggregated or split to satisfy some implementation constraint,
- In other cases, the concept has not been taken into account during data set design, but it can be retrieved by simple operations on represented concepts.

As a conclusion, users who only have metadata description at their disposal, or who are unable to decode the complex technical description, cannot assess whether the information content they need is provided by the data set. That makes the selection of an appropriate data set arduous. The need for data pre-processing is described in the next section

1.3. ISSUES INVOLVED IN PRE-PROCESSING DATA SETS

Data are generally produced by organisations such as mapping agencies that have a limited number of immediate application purposes. So, data sets cannot be customised for every specific need, and they seldom exactly suit the user application requirements. Even if a data set provides the right information content, pre-processing may be a necessary step to adapt data to the application. There can be a problem of data structure (e.g. feature types must be split or aggregated), format, spatial representation (e.g. topologic relationships must be calculated so that roads form a network), spatial referencing (e.g. geographic coordinates must be converted to another system), etc.

Data transformation makes it possible to change data format or geometry, to modify a schema structure, to filter information, etc. But these transformations may be time and effort-consuming for non-expert users. Moreover, these home-made transformations are not referenced by the data provider, which does not guarantee any compliance with differential data sets when the data must be updated [13]. That is why a user catalogue query should be composed not only of an extraction query, but also of a pre-processing order. This customisation step should be supported and referenced by the data provider.
Such a customisation of existing data sets for specific needs has not been widely studied. The UAPE system presented by De Oliveira [14] tries to adapt data and applications to the user platform: it allows the user to design his own database and application, which then should be automatically implemented by a driver on the chosen GIS. Hubert [15] adapts data representation, through a cartographic generalisation process, to the user requirements of visual data display.

To summarise the issues involved here, crucial steps in user access to geographic data include the specification of a data set containing the required information (i.e. the choice of an information content based on metadata and data set technical documentation) and the restructuring of this data set to make it application compliant (i.e. the choice of a representation for the selected information content).

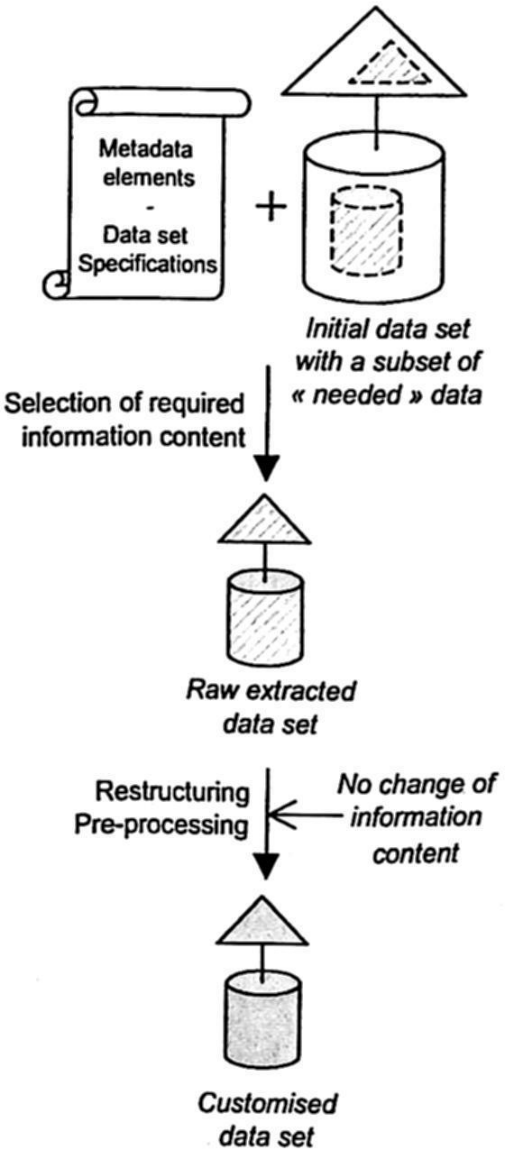
Metadata elements
-
Data set Specifications
+
Initial data set with a subset of « needed » data
Selection of required information content
Raw extracted data set
Restructuring Pre-processing
No change of information content
Customised data set

Figure 4. How to process customised data set specification

These steps are represented in Figure 4. They can hardly be achieved by end users because of the lack of available and explicit information describing data sets, and because of the complexity of the information structure. However, technical documentation of geographic databases and simple operations could help assess and improve data sets fitness for use. In our research, we intend to use these elements to build a system for customised data sets specification.

2. APPROACH

We aim at designing a system enabling users to order customised data sets, i.e. to assess existing data set content, to select within existing representation elements the information content the most adapted to their need, to restructure it, and to specify some additional pre-processing. The outputs of the system are orders to be placed to the data provider:

- An order for data extraction corresponding to the needed information content,
- An order specifying some pre-processing to be carried out on extracted data so that the data schema and format are consistent with the user application.

The two main elements of our approach are:

- A data set description model which links the user needed geographic features (e.g. roads) to implemented tables (e.g. road_section), possibly with data structure transformations (e.g. aggregation of road_section objects).
- An interface to support interaction with the user. It is an essential part of the system. A lot of work must be completed to display the rich and complex information of the data set description model in a simple form. Since the specification of customised data sets is an interactive process that requires user-system negotiation, a dialogue process controlled by a dialogue manager must be defined. Among other things, a dictionary will be required to match user key-words and system known concepts.

The steps of the specification process are presented in section 2.1. Section 2.2 focuses on the data set description model.

1.3. THE CUSTOMISED DATA SET SPECIFICATION PROCESSING

Figure 5 represents the successive steps that will be proposed by our system for interactive data set specification. These steps are then briefly described. Some of them rely on user-system interaction, others rely on the system only.

To illustrate sections 2.1 and 2.2, we take the example of a user who needs the road network, the rivers and their crossing.

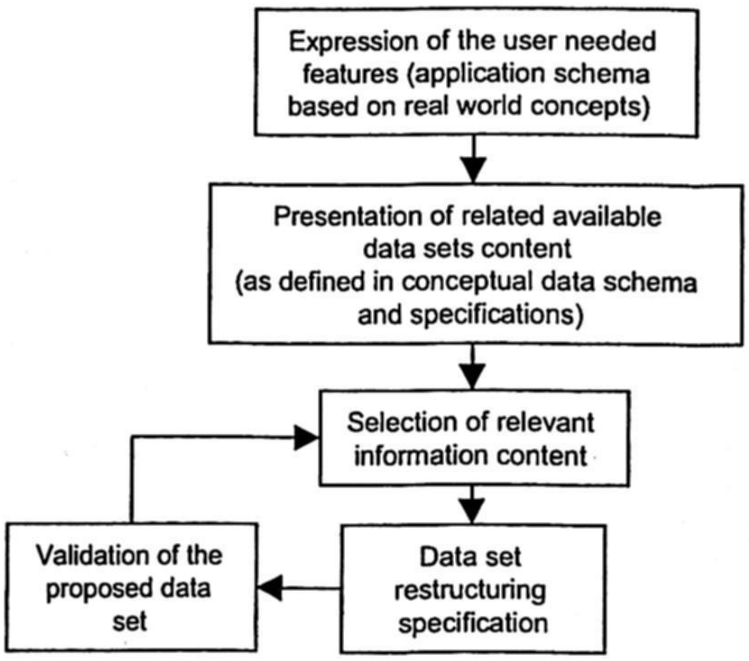
Expression of the user needed features (application schema based on real world concepts)
Presentation of related available data sets content (as defined in conceptual data schema and specifications)
Selection of relevant information content
Data set restructuring specification
Validation of the proposed data set

Figure 5. Main steps in the customised data set specification process

Expression of user needed features.
The first step is the expression of user needed features. The information required by the user for his application can be represented in an application schema. ISO provides guidelines for application schema definition [16]. Since we suppose that non-expert users need a simple way to express their need, we situate the user application schema at a less complex and detailed level: we will use semantic networks to propose features and define an application schema. An example of user application schema is shown in Figure 6.

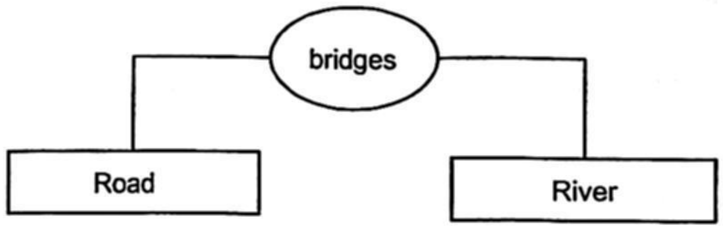

Figure 6. The user application schema

Presentation of data set content.
The second step is the presentation of available data sets information content. The system identifies existing representation elements of the conceptual data schema corresponding to the needed features, and presents them. Real data samples are also displayed, and all useful information available from metadata and data set specifications is provided so that the user can precisely assess the information content of the proposed representation elements. For the specific GeoBase data set, the proposed representation elements are those shown in Figure 7. If the system "knows" several data products likely to provide the adequate information content, it must present them distinctly and allow comparisons.

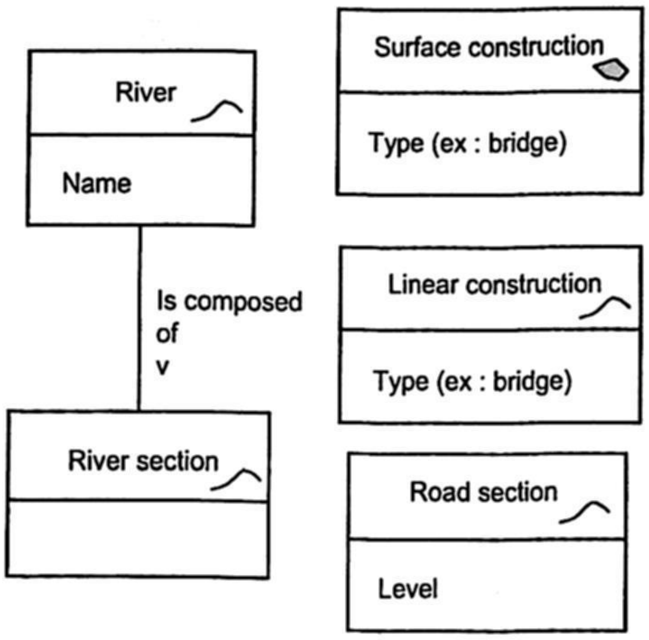

Figure 7. Representation elements of the GeoBase conceptual data schema, likely to correspond to the user application schema elements.

Selection of relevant information content.
During the third step, the user selects the most relevant representation elements. For example, after a look at the GeoBase content specifications and data samples, he selects the elements "river", "road section" and "linear construction" from the displayed representation elements (Figure 7).
This selection step should not be seen as a query definition: the goal is not to pick some objects from existing representation elements, but rather to define new representation elements. It can be seen as a view definition process, with this difference that the defined view is going to be restructured and extracted and will result in a new independent data set.
The representation elements must be chosen within a single data schema, even if several products have been presented during the previous step. Indeed, the system is not meant to integrate heterogeneous information, but to adapt and customise existing products.

Data set restructuring and pre-processing.
The fourth step is the restructuring and pre-processing of the selected data set. The system has to propose a customised data set as close as possible to the application schema expressed by the user, e.g. that explicitly represents the concepts that appear in the application schema. This can imply the need to restructure the data set to derive new components. For this step, the system can appeal to specific operations (i.e. class fusion, filtering, aggregation, etc.). This step relies on the data set description model and is further detailed in the next section.
Some simpler pre-processing operations such as changes of data format or coordinate system are necessary to fully adapt the data set to user requirements. They cannot be carried out on feature types, but directly on data. They can be ordered by the user and registered by the system, so as to be executed later, after the data extraction phase. On the other hand, treatments requiring complex tuning such as cartographic symbolisation or generalisation are not taken into account.

Validation: checking application compliance.
The fifth step is a validation step. It checks that the specified customised data set is accepted by the user and consistent with his planned applications. Indeed, as explained in section 1, each application requires specific data properties, which will be defined in a rule base. As every specific application requirement cannot be described, we limit this validation to common operations concerning measurement and selection, locating and addressing. Most complex applications are composed of such basic operations.
If the constraints induced by the user's planned activities are not fully respected by the specified custom-made data set, the system has to point this out. Previous steps of selection, restructuring and pre-processing have to be corrected or renewed.

1.1. THE DATA SET DESCRIPTION MODEL

In section 1, the gap between the real world features required by the user and the available descriptions of data set content was pointed out. Several modelling levels were distinguished. In this section, we present our draft data set description model that bridges the levels of needed features (real world), representation elements (conceptual data schema) and stored tables (logical data schema) together.

The problem analysis points out the need for an enriched data schema, as shown in Figure 8, with an intermediary conceptual level between the user application schema and the initial conceptual data schema.

Indeed, the initial conceptual data schema can hardly be connected to user needed features for several reasons that can be recalled here:

- It provides dense information. Depending on the level of detail, one real world feature may be related to several representation elements whose names are not always explicit. It is all the more dense in that it provides detailed information related to acquisition, representation and storage constraints.
- It does not describe the whole information content. The features that are implicitly represented in the data set are not taken into account.

To be easily exploited for user needs, this initial conceptual data schema should be simplified when too dense information is given, and enriched with concepts implicitly contained in data. This is the role of the enriched conceptual data schema.

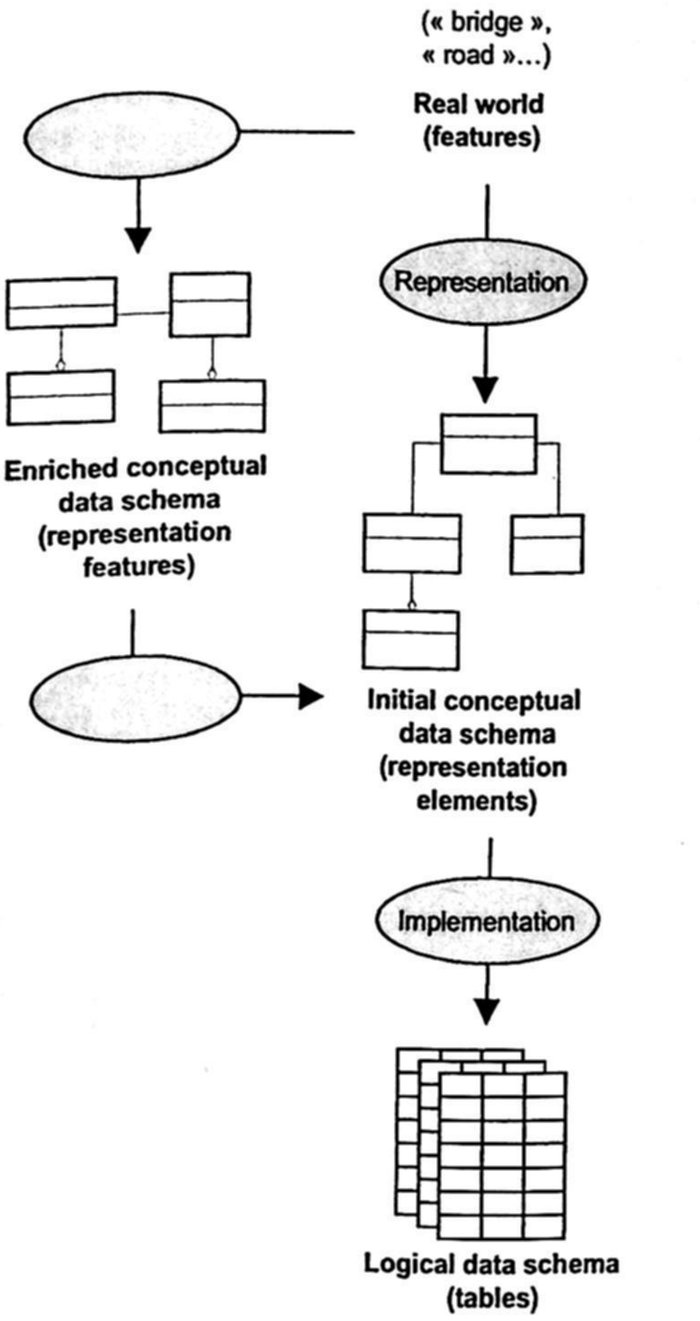

Figure 8. Use of an enriched conceptual data schema

In our data set description model, the relationship linking representation features of the enriched conceptual data schema to existing representation elements of the initial conceptual data schema is a "yield" relationship. This relationship must be supported by simple schema transformation operations, as shown in Figure 9.
Real world features mentioned in the user application schema are related to representation features of the enriched conceptual data schema through relationships of "possible representation".
To distinguish abstraction levels, we borrow the concept of *representation stamp* [17] from the MADS formalism for multiple representation data:

- User needed features carry the "usr" stamp,
- Representation elements of the initial conceptual data schema have the stamp "ini"
- Representation features derived from the initial conceptual schema by transformation operations carry the stamp "implicit" or "imp".

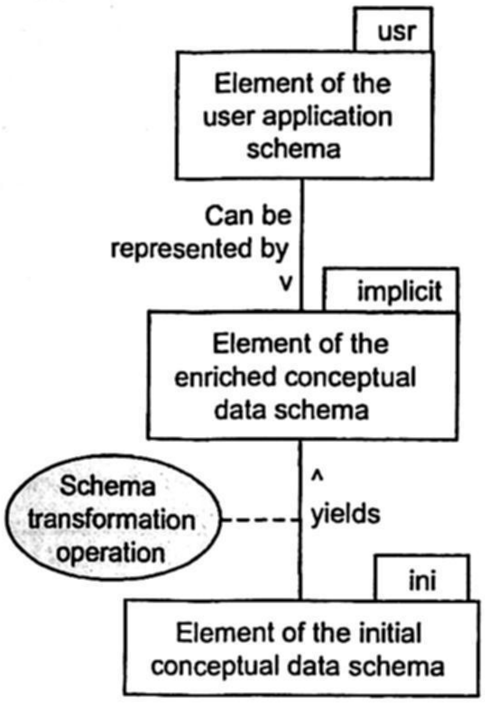
usr
Element of the user application schema
Can be represented by
v
implicit
Element of the enriched conceptual data schema
Schema transformation operation
^
yields
ini
Element of the initial conceptual data schema

Figure 9. The "implicit representation " and "possible representation" relations

Some of the transformation operations can be automatically proposed by the system, such as the aggregation of the existing "Road section" element to create the desired "Road" feature. Others must be requested by the user.
In our example, the "Road" feature and the "bridging" relationship have to be built through operations shown in figure 9.

The features of the proposed customised data set maintain some properties of the initial conceptual data schema elements, which enable them to support user applications or not. The proposed data set in Figure 9 can be used to select roads and rivers concerned by a bridging relation. But location of bridging points or measure of the bridge length are not possible, whereas the initial data set contains the required information. If such operations are required by the user application, another solution must be found that derives a "Bridge" feature with a geometry.

3. CONCLUSION

This paper intends to propose an approach to deal with the problem of user access to geographic products suiting their need. We aim at building a system designed to specify a customised data set. The core elements of our data set description model have been presented. This model mainly relies on an enrichment of available data sets: it must describe precisely the properties of available representations elements. It finally must include the simple data manipulation operations that can be used to derive a data set explicitly representing the user needed features, in a way adapted to its intended use.

This approach makes the data provider in charge of the process of data restructuring and representation adaptation. By the way, the data provider keeps a trace of the data set modification and is able to assure future data set maintenance.

In this project, the source data sets will be limited to existing IGN data sets, and the described content will probably be limited to a single theme like water bodies. Even if it dramatically reduces the area of possible custom-made data sets, it will enable us to explore a new type of access to geographic data. This is likely to improve usability of existing products without modifying their structure or their content.

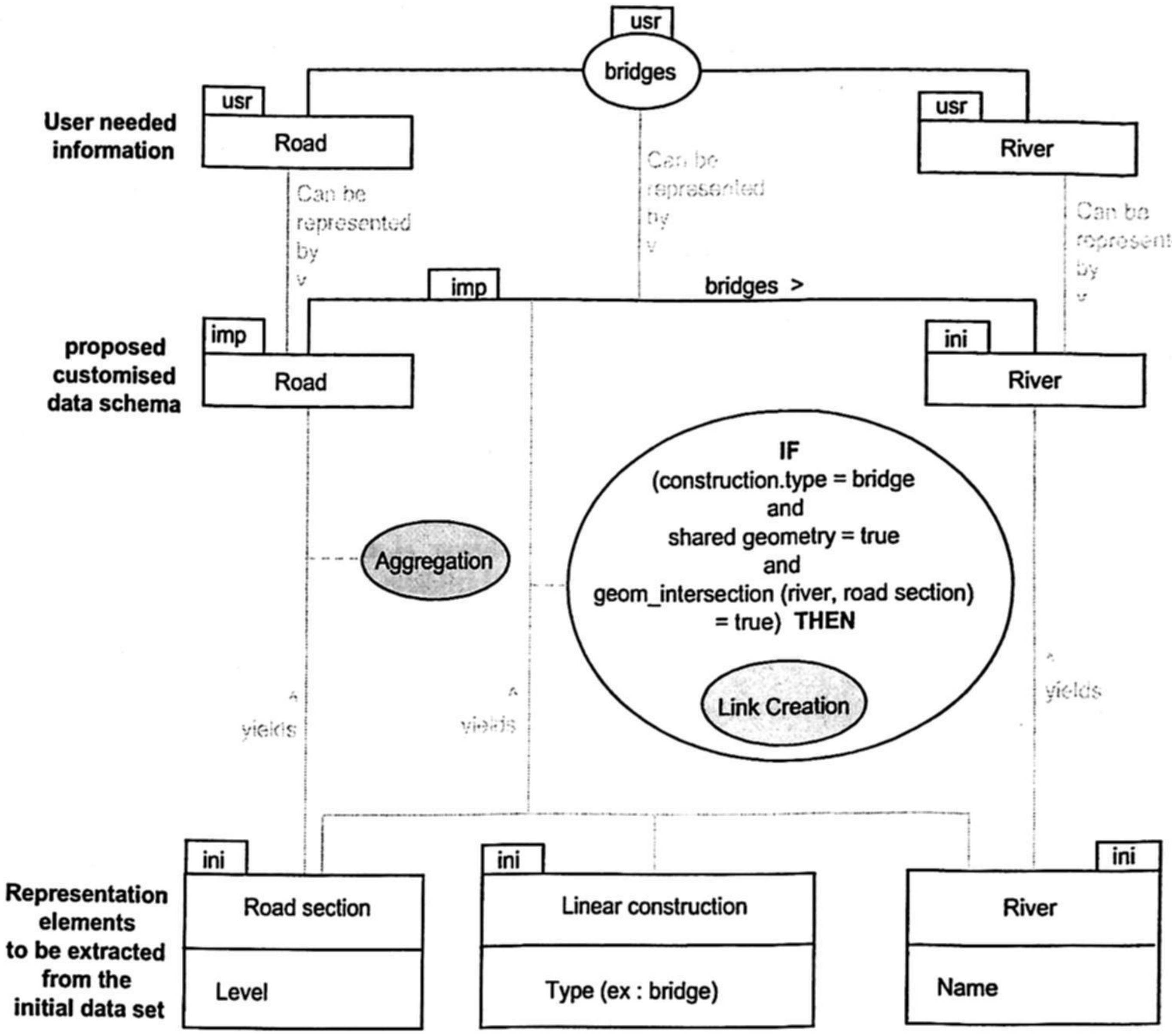

Figure 9. Data set restructuring. Links of implicit and possible representation are shown by pale arrows.

REFERENCES

[1] A.U. Frank & W. Kuhn, Specification Languages for Open GIS, *Revue Internationale de Géomatique, 9(2)*, 1999, 135-152.

[2] B. Bucher, A Model to Store and Reuse Geographic Application Patterns, *Proc 4th Agile Conf on Geographic Information Science*, Brno, Czech Republic, 2001, 289-295.

[3] A.U. Frank & M. Jahn, How to Sell the Same Data to Different Users at Different Prices, *Proc 6th Agile Conf on Geographic Information Science*, Lyon, France, 2003, 357-360.

[4] B.Vasseur, R. Devillers & R. Jeansoulin, Ontological approach of the fitness for use of geospatial datasets, *Proc 6th Agile Conf on Geographic Information Science*, Lyon, France, 2003, 497-504.

[5] GSDI Technical Working Group, Developing Spatial Data Infrastructures : the SDI CookBook, v1.1, Douglas Nebert (Ed), 2001, on line: http://www.gsdi.org/pubs/cookbook/cookbook0515.pdf

[6] P. Ahonen-Rainio, Description of the Content of geographic datasets, *Proc 8th SCANGIS Conf on Geographic Information Science*, As, Norway, 2001.

[7] ISO TC211, ISO 19115 Geographic Information - metadata, international standard, 2003

[8] G. Gardarin, *Bases de données*. (Ed Eyrolles, 2000).

[9] F.T. Fonsesca & M.J. Egenhofer,. Ontology Driven Geographic Information Systems, *Proc 7th ACM Symposium on Advances in Geographic Information Systems*, Kansas City, USA, 1999, 14-19.

[10] S. Mustière, N. Gesbert & D. Sheeren, Formal Model for the Specifications of Geographic Databases, *Proc GEOPRO International Workshop on Semantic Processing of Spatial Data*, Mexico City, Mexico, 2003.

[11] C. Rüther., W. Kuhn.& Y. Bishr., An Algebraic Description of a Common Ontology for ATKIS and GDF, *Proc 3rd AGILE Conference on geographic information science*, Helsinki, Finland, 2000.

[12] G. Goder, Représentation comparée de schémas et spécifications de contenu, Master thesis, 2003.

[13] T. Badard, On the Automatic Retrieval of Updates in Geographic Databases Based on Geographic Data Matching Tools, *Proc 19th International Cartographic Conference ICA/ACI*, Ottawa, Canada, 1999, 47-56.

[14] J.L. De Oliveira, F. Pires & C.B Medeiros, An Environment for Modeling and Design of Geographic Applications. *GeoInformatica, 1*, 1997, 29-58.

[15] F. Hubert, B. Bucher, S. Balley & A. Ruas, A User Interface to Specify the Parameterisation of GIS Function: The Example of Building Generalisation Parameterisation, *Proc EuroSDR Workshop on Visualisation and Rendering*, Enshede, Netherlands, 2003.

[16] ISO TC211, ISO/DIS 19109 Geographic Information - Rules for application schema, draft international standard, 2003

[17] C. Vangenot, C. Parent & S. Spaccapietra, Modeling and Manipulating Multiple Representations of Spatial Data, *Proc SDH Conf on Advances in Spatial Data Handling*, Ottawa, Canada, 2002, 81-93.