Ontology Integration for Statistical Information

Wilfried Grossmann and Markus Moschner

WG Data Analysis and Computing, Dept. of Computer Science,
University of Vienna, A–1010 Vienna, Austria

Abstract. Metadata may be used for convenient handling of statistical information. Thus some metadata standards have emerged as guiding lines for information processing within statistical information systems. Another development stems from documentation development for data archives where metadata requirements of researchers in social sciences are considered. Different metadata standards have emerged out of these streams of science. Basic ideas on integration and translation between such different lines of development are given. Hereby principles of ontology engineering play a key role as starting point.

1 Introduction

Statistical information plays a central role in many business and economic decisions. The term information means here that we have to consider the data itself as well as descriptions of the data, so called metadata, which are necessary for obtaining the information. Hence, it is not surprising that metadata play a key role in statistical information systems for a long time. The earliest reference is Sundgren ([13]), who introduced the concept of metadata in statistics for a so called infological approach towards statistical information. This approach has been developed further in many ways by a number of researchers as well as statistical agencies and has lead to a number of metadata standards. Due to the fact that a lot of statistical information is contained in highly aggregated data, represented nowadays usually in data warehouses, one line of development focussed on metadata standards for such type of data (for example the SDDS-standard [11]). A second stream of development based on ideas from documentation for data archives considered mainly the metadata requirements for scientific researchers in social sciences and economics. These efforts have resulted in different metadata standards, probably the best known example is the DDI standard [3], which is a substantive expansion of the Dublin Core metadata [2]. A well known software tool based on these ideas is NESSTAR [8]. A further development concentrated on proper metadata representation for value domains of the attributes of interest, resulting in the so called Neuchatel Terminology [9] for classifications.

Due to the different starting points of these approaches it is rather cumbersome to integrate data in cases where the definition of the data schemes is based on such different documentation schemes. Following the developments of

R. Moreno Díaz et al. (Eds.): EUROCAST 2005, LNCS 3643, pp. 63–68, 2005.

intelligent information processing in recent years the field of statistical information processing has seen a number of efforts to develop the idea of metadata further into the direction of ontology (see for example Sowa [12]). In fact, statistical metadata practice includes to a far extent information needed for ontology - engineering. Probably the most important contribution in this direction was made by the METANET [5,7] project, a thematic network within the fifth EU-research framework. The approach tries to fulfil the requirements of the ontology definition of Gruber [6] (”ontology” is a specification of a conceptualization) by formalizing the statistical data paradigm, taking into account the representational, operational and functional semantics of statistical information.

Starting points are basic statistical objects like dataset, statistical population, or statistical variable, which constitute the categories for the ontology. For all these categories a unified description framework was developed, which is called the view facet of statistical categories. The following four views were distinguished:

- The conceptual category view represents the subject-matter definition of any category instance and builds the bridge to reality. Validity of the definition of the subject matter concept gets usually restricted by temporal and geospatial constraints.
- The statistical (methodological) category view describes the statistical properties of the category instance by using a number of parameters, which have to be defined in such way that specific properties of the different categories are taken into account.
- The data management category view is geared towards machine supported manipulation storage and retrieval of the category instance data.
- The administrative category view addresses management and book-keeping of the structures.

Based on these view facets a representation scheme can be defined, which seems to be sufficient for operational purposes. A first sketch of such a model was presented in Denk et al. [4]. In this paper we present first and fundamental ideas for using this framework for mapping different metadata standards.

2 Methodologies

Though scientists speak different languages there is still communication and consent on subjects in question possible. Different metadata standards are only partially the consequence of unintelligibility on research subjects and basic formulations. There is still enough communication about such differences possible. Here we concentrate only on such standards which basic principles and formulations can be systematically treated by humans from a bird’s eye view. Only from such a unifying treatment formalizations are tackled, as there are knowledge representations, order sorted algebras, data types, mathematical approach(es) and statistical notions (units, population and statistical variable). In the following the main issues are given in systematic order.

2.1 Fundamental Concepts

All metadata standards can get formulated by a common basis of fundamental concepts (foundational basis). There may be different ways to do this, yet one particular way is chosen. Such fundamental concepts can be arranged like in formal ontologies. That means that there are atomar notions within a partial order where such an order means something like sub- or superconcept. Set-like operations (join, meet and complement) get induced by such an "order". The used vocabulary comprises attribute-like properties and restricted quantifications like in description logic ([10]). Herewith the basis for the intended fundamental concepts get formulated such that differences and relevant properties get included. It will not be realistic to aim at a world knowledge nor will some technical issues (following below) be settled. For special cases even different ontological approaches might be chosen.

The aforementioned basic statistical objects (population, units and variables) have to fit into the chosen conceptualization.
Concepts are described by basic notions and binary relations between these notions. Hence a set-like formulation for a fundamental structure lies at hand $K =< BNotion, BRel, Subsump, Meet, Join, Compl, Null >$, a pool Var of Variables (there might be more than one sort), properties of elements of K analogous to predicates (roles) and quantifiers in description logics([10]).

2.2 Order Sorted Algebras and Data Types

Til now only abstract concepts have been considered. Handling of values as numbers is one part of statistical information processing. Analogous to programming languages specification of data types with concrete value domains is mandatory. There are data types like numbers and strings which do not share much or even nothing (from a conceptual viewpoint). Furthermore some data types form (nontrivial) order sorted algebras. One example are natural, integer, rational and real numbers where operations are also extended. Intervals will play a prominent role with respect to numbers. The possibility of (domain) restrictions need the concept of attributes - monadic predicates which may be used as generators for new sorts. One may distinct between sub- and supersorts with respect to the partial order of the conceptual basis and sorts associated with subset properties.

A datatype $D =< DDom, DPred >$ represents a domain with predicates defined on it (operations have to be formulated as some sort of equations - what seems rather naturally for most cases). For each datatype we have an instantiation (or: intended interpretation) DInt which maps D into some grammatical structure DStr.

2.3 Formalization of Mathematical Concepts

There is a strong relation between sorts of formal mathematical content and data types. Formalization brings a large body of ordered sorts. It is not always necessary to have a correspondence between formal sorts and ontological concepts.

Formalization gives (semi-)automatic processing of mathematical statements, thus only basic or important mathematical notions need a correspondence to ontological concepts. Since there is (and will be) a versatile variety for this discipline([1]) reinventions should be avoided.

2.4 Statistical Variables

A statistical variable SV is a (partial) mapping from K into instantiations $DsStr$ of datatype domains.

There is not only ontological meaning behind. Statistical variables give also concrete values for abstract notions. In that sense they play a central role in being a bridge between abstract notions and value domains: Thus it is legitimate to see them as interpretations of metadata standards into the available order sorted algebras. The statistical notions of unit and population get hereby a determination with concrete values.

2.5 Specification for Data Repositories

Hereto belong merely technical specifications like data base or storage organization. Detailed inspection of data management issues is beyond the scope of this paper.

2.6 Metadata Formalization

Concrete metadata standards get reformulated pertaining to the fundamental concepts. The original formulation of metadata standards gets simply mapped onto ontological notions whereas totality is not compulsory. What is likely to appear is that especially for technical notions such a standard might be more fine grained than the used ontology. Either start again with an enhanced foundational basis or construct an according coarse mapping. Such a mapping does not mean to forget such more fine grained standard notions- yet only its aggregation with certain ontological notions. That is not only a matter of lowering the work for ontology construction but sometimes one does not need or has no justification for differentiation between some technical terms. In matters of the ontology these are too near related.

It is to expect that such metadata standards are like taxonomies. At least there has to be a set of TC of concepts which is partially ordered.

If description logics are used for specification then a metadata standard (or a substantial part of it) gets modelled as a T-Box.

2.7 Morphisms Between Metadata Formalizations

When the foundations are established translations between metadata standards may be tackled. Since ontological foundations shall be taken into account such morphisms are not simply between taxonomical standards itself. Yet these are

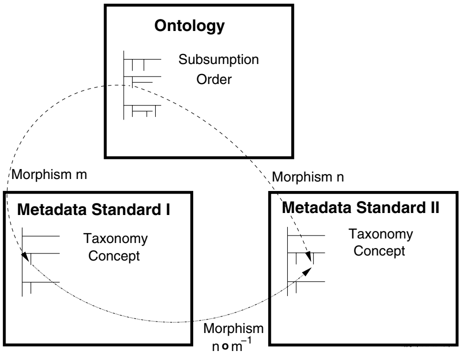

Fig. 1. An arbitrary selection of a translation between metadata standards

between the relations of the fundamental concepts and metadata standards. These morphisms need to be isotone in the sense that the partial order of concepts with respect to subsumption in the foundational basis has to be kept for mapped pairs.

Such a morphism M can be seen technically as subrelation of the support of K and TC. Hereby isotony applies as a basic constraint, yet further may be useful.

The figure exemplifies a simple subcase which would be only a part of a full formalization. Especially for the "metadata standard II" there are two choices. Here experts or other knowledge is necessary for a decision. The morphism between the standards is effectively constructed by use of the fundamental concepts. If done by hand humans play the role of the fundamental concepts. Thus some explication of this activity is demanded here — a task that is too often underestimated.

3 Synopsis

The points of the last section correspond in a certain way to the view facets of statistical categories:

2.1 represents the conceptual category view where matters of knowledge structures are addressed. 2.2-2.4,2.6 and 2.7 comprise statistical approaches as well as concrete object properties. That makes them counterparts of statistical categories. 2.5 resembles the data management and administrative category view.

References

1. Caprotti, O., Dewar, M., Turi, D.: Mathematical Service Matching Using Description Logic and OWL;
   In: Asperti, A., Bancerek, G., Trybulec, A. (eds.): Proc. 3rd Int. Conf. on Mathematical Knowledge Management (MKM 2004), Białowieża, Poland (2004), p. 73–87.
2. DCMI Usage Board, DCMI Metadata Terms, 2003, Dublin Core Metadata Initiative (DCMI), `http://dublincore.org/documents/2003/03/04/dcmi-terms`

3. DDI Data Documentation Initiative — A Project of the Social Science Community–Codebook. http://www.icpsr.umich.edu/DDI/codebook/index.html
4. Denk, M. Froeschl, K.A., Grossmann, W.: Statistical Composites: A Transformation Bound Representation of Statistical Datasets.
In: Kennedy, J. (ed.): Proc. 14th Int. Conf. on Scientific and Statistical Database Management, IEEE Los Alamitos, California/USA (2002), p. 217 - 226.
5. Froeschl, K.A., Grossmann, W., delVecchio, V.: The Concept of Statistical Metadata.
MetaNet (IST–1999–29093) Work Group 2, Deliverable 5 (2003).
6. Gruber, T. R.: A Translation Approach to Portable Ontologies.
Knowledge Acquisition, 5(1993), p. 199-220.
7. The *MetaNet* project, http://www.epros.ed.ac.uk/metanet.
8. NESSTAR — Networked Social Science Tools and Resources,
http://www.nesstar.org.
9. Neterstrøm, S., et. al.: Neuchâtel Terminology — Classification Data Types, Object Types and Their Attributes. Version 2.0;
The Neuchâtel Group, Neuchâtel (2002).
10. Pan, J.Z., Horrocks, I.: Web Ontology Reasoning with Datatype Groups;
In: Fensel, D. et al.(eds.): Proc. of ISWC 2003, LNCS 2870, Springer–Verlag, Berlin (2003).
11. SDDS — Special Data Dissemination Standard, Dissemination Standards Bullentin Bord, 0.http://dsbb.imf.org/Applications/web/sddshome/#metadata.
12. Sowa, J. F.: Ontology, Metadata, and Semiotics.
In: Ganter, B., Mineau, G.W. (eds.): Conceptual Structures: Logical, Linguistics, and Computational Issues. LNAI 1867, Springer–Verlag, Berlin (2000).
13. Sundgren, B.: An Infological Approach to Data Bases;
Urval, nr. 7, National Central Bureau of Statistics (Stockholm, Sweden), 1973.