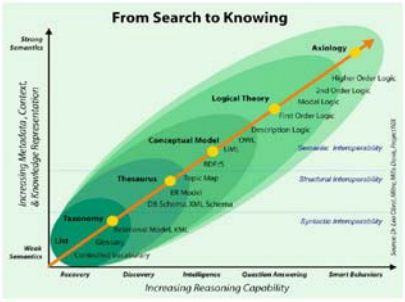
Using the National Education Data Model from a Data Mart Development Perspective

By Jason Wrage July 2008

Introduction



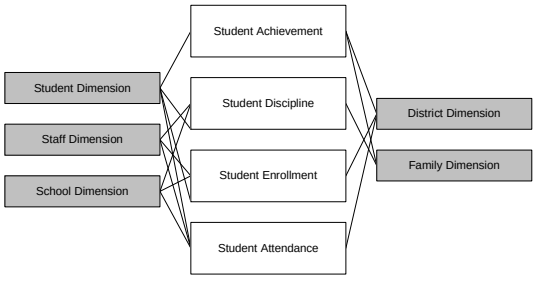
When Vince Paredes, SIFA's Data Model Architect, presented his work on the National Education Data Model at the SIF Association board retreat in Columbus, OH in November 2007 I knew very little about the semantic web and related technologies. During the board retreat Mr. Paredes presented a slide with a graphic visually reminiscent of a football that describes a continuum of interoperability ranging from syntactic on the low end, semantic on the high end, with structural somewhere near the middle. That slide, and the accompanying presentation, caused me to become immediately interested in the potential of semantic technologies. At the time I was unsure how, or why, we needed another data model, but I did not let that get in the way of what it might be possible to achieve using a semantic approach to modeling data. Since the board retreat in November I have studied some on the topic, and would elevate my self-evaluated knowledge level from knowing "very little" to knowing "some" and desiring to know "much more!" This paper discusses the process we use to develop data warehouse and analytics applications, and how the National Education Data Model has intersected with that work.

Project Background

Our project scenario involves a Local Education Agency in Central Illinois that has very specific, and somewhat atypical, reporting needs. These needs were created based on the outcome of a class action lawsuit that requires the district to report various data on specifically defined racial groups (these groups are not the same as the district's AYP subgroups). Generation of these reports, which are referred to as the consent decree reports, takes the district's data and technology teams about 400 hours per quarter to produce using a manual process. After completing an RFP process Integrity was selected as the solution provider.

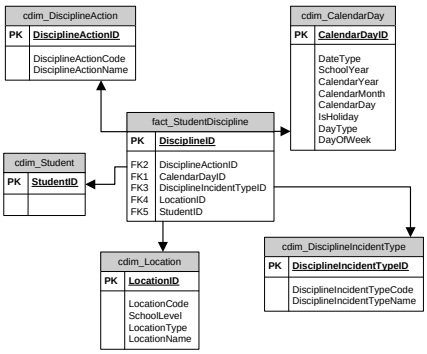
This project started in the spring of 2008 and involves two major deliverables. The first deliverable is a SIF implementation. In this scenario, SIF will provide both application interoperability and first stage extract-transform-load (ETL) functionality for a Data Warehouse that will satisfy the district's reporting needs. A high level depiction of the solution's data flow is shown here.

The second major deliverable is the data warehouse. We are using the Ralph Kimball [KIMBALL] paradigm of building the data warehouse incrementally by first constructing data marts, roughly per business process, and then integrating the marts using conformed dimensions. This methodology enables constructing the data warehouse in reasonably sized chunks. Kimball describes this design as a data warehouse bus since the various marts connect along relationships to a common set of dimensions. Some of the data marts that will be included in the solution include: student enrollment, student attendance, student programs, staff development, and discipline. The remainder of this paper will focus on the discipline data mart that was developed for this solution.

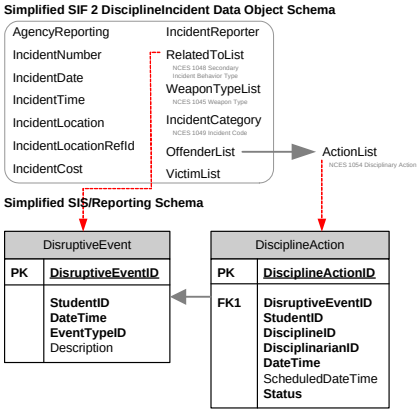


Data Analysis Processes: Top Down and Bottom Up

In order to apply the aforementioned data warehouse construction methodology, we began with a requirements gathering exercise. During that exercise we looked in detail at the existing processes and data used to generate the consent decree reports. This data analysis work was accomplished by reviewing the current set of reports, reviewing the underlying reporting databases the district had created by manually extracting and importing data from their student information system (SIS) into a desktop database application, and interviewing members of their data team. Initially a mind map tool was used to document requirements. This was transformed into a data dictionary which, finally, was used to develop our data mart design. A simplified version of the discipline data mart design can be seen at the right. In total, the deliverables for this portion of the exercise comprise the top down requirements.



Parallel with the data warehouse requirements gathering, we worked on mapping the native SIS data structures to SIF 2 [SIFIS22]. This mapping needed to take place in order to build a custom SIF 2 agent for the SIS. Building the SIF 2 agent was necessary because the SIS's native agent does not, and probably never will, support SIF 2. SIF 2 is required for the first stage ETL because its data model contains discipline data, the DisciplineIncident data object, which was not present in SIF 1.5R1 [SIFIS15R1]. The deliverables for this portion of the project make up the bottom up requirements.



At this stage of the process we noted a major difference between SIF and two other artifacts guiding our design: the SIS discipline data structures and the reporting requirements. The latter two structures are consistent in how they distinguish the action of the student from the punishment assigned by the disciplinarian. The SIF DisciplineIncident data structure essentially creates a single object that encapsulates and relates both of these entities. One of the consistent challenges that SIF deals with is the design and creation of data structures for interoperability. Myriad data models in the universe of operational data systems often leave us with compromises in the way we design Data Objects. The spectrum of design choices ranges from small, highly normalized objects that resemble source-system database structures, to very large, monolithic objects that attempt to combine many related concepts into a single object instance. Falling toward the monolithic end of this spectrum, the SIF 2 DisciplineIncident Data Object makes it relatively easy for an application to send a single request and receive in response relevant discipline data in a single messaging transaction. However, due to the way we need to report this data, the DisciplineIncident design will not work well for the discipline data mart. This is not intended as any type of criticism regarding the design of DisciplineIncident. Again, the object addresses interoperability well, which is SIF's reason's for being, but is not optimal in this case for direct reporting. To solve this problem, the second stage ETL transforms the incoming SIF data into the data mart schema.

National Education Data Model Gap Analysis

Having made our way through the requirements gathering, analysis, and initial design processes of our project a few weeks ago, we began comparing our work with the discipline structures in the National Education Data Model. Several tools are available to put the Data Model to work. A table in the Resources section of this document shows a short list of tools with links and comments. We started the comparison using the National Data Model Browser and by entering a keyword search for the term "discipline." This returned two results: a class named "DisciplineEvent" and an instance (also called an individual or entity) named "disciplineIncident." The description provided for disciplineIncident is "an event classified as warranting discipline action." The description provided for DisciplineEvent is "an event that follows a discipline incident." Initially, our reaction was positive because we saw

in the National Education Data Model the separation of an event requiring a disciplinary response, from the disciplinary response itself. Given the collisions in the use of our terms with the National Data Model terms we decided to look deeper into the structure to confirm the perceived similarities.

The next grain of detail that we analyzed were the "child" nodes of DisciplineEvent and disciplineIncident. At the time of the analysis, the table at the right represents what the data model contained (this has since been enriched based on some of our suggestions). The values "detention," "suspension," and "expulsion" were strong clues that the National Education Data Model's DisciplineEvent class corresponded to our data mart's Discipline Action conformed dimension. To confirm this we sought, but could not find, a relationship between DisciplineEvent and disciplineIncident that would indicate causality.

Child Nodes

disciplineIncident (i)	DisciplineEvent (c)
schedule (a)	detention (i) suspension (i) expulsion (i)

(c)lass, (i)nstance, (a)ttribute

One of the key advantages of an ontology-based data model is the ability to express any type of relationship among objects. Traditional entity-relationship models are extremely constrained in that relationships only indicate cardinality. With XML-based data models it is possible to achieve richer relationships through the use of context and encapsulation. However, this is generally accompanied by a drop in precision versus pure relational models. Using an ontology-based data model enables rich relationships among objects without a loss of precision.

Continuing to seek the relationship that should exist between the discipline objects, we browsed the student instance and discovered two types of relationships being used to connect it to four other objects. It was then apparent that at least some areas of the National Education Data Model are making use of relationships. Left with more questions than answers on the missing discipline relationships, I decided to address the question to Mr. Paredes. The answer was simple: the National Education Data Model is evolutionary. Although a causal relationship between DisciplineEvent and disciplineIncident would seem apparent, it had not yet been suggested by reviewers; therefore, it did not yet exist. At the time of this writing, Mr. Paredes is exploring the best way to implement the relationship between DisciplineEvent and disciplineIncident. We also discussed several new instances for DisciplineEvent that came from our data mart design, which have since been added to the National Education Data Model.

Conclusion

From the process of exploration and the dialog with Mr. Paredes we attained a new affirmation regarding our data mart design. We were able to confirm that our requirements gathering and design made sense but we were also to contribute back to the Data Model. The data model confirmed our logical model but needed more detail.

Almost all of our projects deal with mapping one form of data to another. Using SIF tends to reduce the amount of low level mapping work that must be done in the layers close to operational data systems. Using standardized schemas in the data mart layer also introduces efficiencies to the process. With this particular project, we found significant value in confirming what we thought was a good data mart design for discipline by comparing it to the National Education Data Model. We also learned first hand of the potential for the National Education Data Model to grow and adapt based on real-world influences. As more users and reviewers of the data model provide feedback, it will become more comprehensive and provide significant value to many.

In my personal opinion, homogeneity in data model design should not be the ultimate goal for the National Education Data Model. Like software, data models differ because they need to solve different problems for different people. As semantic technology becomes mainstreamed, it is foreseeable that much of the one-off data mapping work can be reduced; this would seem to be the minimum benefit. Beyond this, semantic technology and the National Education Data Model offer the potential for more dynamic interoperability across systems. With semantically enabled data models and software, systems will be able to negotiate and establish data contracts on demand, without static, labor intensive data mapping exercises. This will enable very dynamic and accurate exchanges of data and information among systems.

Resources

Tools for Exploring the National Education Data Model

Tool	URL	Platforms	Description
Protégé	http://protege.stanford.edu/	Java on Windows, Mac OS, Linux, Unix, Others	Protégé is the main tool that is used to develop the National Education Data Model. It is relatively sophisticated in terms of its capabilities and user interface. Protégé is IDE-like in its appearance and provides support for plugins to extend its core functionality. Although quite powerful, some learning curve is required.
SIF Data Model Browser	http://nces.sifinfo.org/datamodel	Web browser	The Data Model browser hosted by SIF provides the ability to search by keyword, and browse the data model top-down. It is simple and straightforward to use.
SWOOP	http://code.google.com/p/swoop/	Java on Mac OS, Windows	SWOOP seems to lack some of the sophisticates features of Protégé, but the two tools are similar in their basic capabilities. Between the two, SWOOP seems a bit more approachable.
XML Editors	Various	Various	It is also possible to use XML editors and tools to explore and analyze the National Education Data Model since OWL and RDF are XML languages.

Terms

Term	Source	Definition
Class (OWL)	W3COWL	A class defines a group of individuals that belong together because they share some properties. For example, Deborah and Frank are both members of the class Person. Classes can be organized in a specialization hierarchy using subClassOf. There is a built-in most general class named Thing that is the class of all individuals and is a superclass of all OWL classes. There is also a built-in most specific class named Nothing that is the class that has no instances and a subclass of all OWL classes.
Conformed dimension	Kimball	Dimensions are conformed when they are either exactly the same (including the keys) or one is a perfect subset of the other.
Dimension	Kimball	An independent entity in a dimensional model that serves as an entry point or as a mechanism for slicing and dicing the additive measures in the fact table of the dimensional model.
Entity	EDDM	Synonymous with OWL individual and instance.
Individual, Instance (OWL)	W3COWL	Individuals are instances of classes, and properties may be used to relate one individual to another. For example, an individual named Deborah may be described as an instance of the class Person and the property hasEmployer may be used to relate the individual Deborah to the individual StanfordUniversity.
Ontology	WPONTOLOGY	In both computer science and information science, an ontology is a formal representation of a set of concepts within a domain and the relationships between those concepts. It is used to reason about the properties of that domain, and may be used to define the domain.
OWL (Web Ontology Language)	W3COWL	The OWL Web Ontology Language is designed for use by applications that need to process the content of information instead of just presenting information to humans. OWL facilitates greater machine interpretability of Web content than that supported by XML, RDF, and RDF Schema (RDF-S) by providing additional vocabulary along with a formal semantics. OWL has three increasingly-expressive sublanguages: OWL Lite, OWL DL, and OWL Full.
RDF , RDFS	W3CRDF	The Resource Description Framework (RDF) is a general-purpose language for representing information in the Web. This specification describes how to use RDF to describe RDF vocabularies. This specification defines a vocabulary for this purpose and defines other built-in RDF vocabulary initially specified in the RDF Model and Syntax Specification.
Semantic Web	W3CSW	The Semantic Web provides a common framework that allows data to be shared and reused across application, enterprise, and community boundaries. It is a collaborative effort led by W3C with participation from a large number of researchers and industrial partners. It is based on the Resource Description Framework (RDF).

Sources

EDDM. Schools Interoperability Framework Association (SIFA). Data Model Browser. V1p16. < http://nces.sifinfo.org/datamodel/Default.aspx >

KIMBALL. Kimball, Ross. The Data Warehouse Toolkit (Second Edition). John Wiley & Sons, 2002.

SIFIS15R1. Schools Interoperability Framework Association (SIFA). SIF Implementation Specification Version 1.5R1. 11 October 2004. < http://specification.sifinfo.org/Implementation/1.5r1 >.

SIFIS22. Schools Interoperability Framework Association (SIFA). SIF Implementation Specification Version 2.2. 17 March 2008. < http://specification.sifinfo.org/Implementation/2.2 >.

W3COWL. World Wide Web Consortium. OWL Web Ontology Language Overview. 10 February 2004. < http://www.w3.org/TR/owl-features >.

W3CRDF. World Wide Web Consortium. RDF Vocabulary Description Language 1.0: RDF Schema. 10 February 2004. < http://www.w3.org/TR/rdf-schema >.

W3CSW. World Wide Web Consortium. W3C Semantic Web Activity. < http://www.w3.org/2001/sw >.

WPONTOLOGY. Wikipedia. Ontology. 19 July 2008.
< http://en.wikipedia.org/wiki/Ontology_(computer_science) >.