W3C Workshop on Data and Services Integration

Position Paper

Cory Casanave, Model Driven Solutions

and

Sjir Nijssen, PNA Group

Background and interest in workshop

Model Driven Solutions (http://www.modeldriven.com)

Model Driven Solutions provides solutions to our customers based on well-defined models and leading-edge standards based implementations. Our customers include U.S. Governmental organizations as well as large commercial enterprises. A major problem faced by our clients is data federation – data is flowing from open, commercial and internal sources and this data provides a key element to effectively collaborating, using existing assets and planning through analytics. MDS Helps clients solve these problems with architecture and open source implementations. MDS has contributed to international standards for over 20 years in pursuit of more effective and open software solutions.

While there are advanced tools and techniques that can be applied to the problem the issues with incompatibility in the semantics, terminology, structure and representation of data from independently conceived sources makes data federation, and all the related benefits of collaboration, reuse and analytics, expensive and unreliable. Techniques and technologies are available to make a substantial impact on solving this data problem, yet these are not standard, well supported or mainstream. MDS is interested in industrial strength solutions to the data problem.

PNA Group

PNA Group provides conceptual modeling solutions and associated software to its customers.

PNA has a rich pedigree in modeling vision and advancements. Employees of PNA Group have been involved in development of conceptual modeling ever since the inception of the IFIP WG 2.6 in 1974. A noteworthy example is ISO TR9007, Concepts and Terminology for the Conceptual Schema and the Information Base (1987). The CTO of PNA Group, Prof. Nijssen, is the originator of NIAM, which served as the base for development into ORM.

The latest offering of PNA Group is CogNIAM (Cognition enhanced NIAM), which integrates all relevant aspects of processes, data, business rules, semantics and communication engineering in a single, consistent model providing these different viewpoints. CogNIAM includes a very extensive protocol on how to engineer such a Conceptual Domain Model (CDM) and the associated transformations to standards like BPMN, SBVR, UML, OWL and XSD. Combined with the supporting repository software, this approach makes for sustainable and maintainable CDM's, even for large and complex environments.

Clients of PNA Group include large financial, technical and governmental organizations that apply Conceptual Domain Modeling as an approach to specify the "WHAT" question of business to the 100 % completeness level, before such specifications are implemented through one of various business-driven choices (e.g. implemented in standard software, custom development or off-shoring).

PNA Group has been commissioned by a major international organization to provide the first version of a SIMF-like software repository.

Selected Related Work

Customer Data Federation

In existing projects with clients MDS is actively engaged in solving the data problem for both internal and external process, organization and systems integration. We employ both "Model Driven Architecture" (MDA) and "Semantic Web" techniques and technologies within these projects.

OMG "SIMF" and Architecture Ecosystem

The "Object Management Group" (OMG) is the premiere modeling standards organization supporting such standards as UML, BPMN, SoaML, MOF and ODM. The authors have been active within the OMG to attempt to solve the problems of model federation as well as data federation. The latest effort "Semantic Information Modeling for Federation" (SIMF) directly applies to the workshop goals. The SIMF approach is further elaborated, below.

U.S. NIEM

The U.S. Government sponsors the "National Information Exchange Model" (http://www.niem.gov), a controlled vocabulary represented in XML Schema to assist in information sharing. MDS has helped author and is currently participating with the NIEM community to respond to an RFP for modeling NIEM in UML and thus providing for a higher level of abstraction and support for multiple technologies from the same model using "Model Driven Architecture" techniques.

SIMF Approach

The approach presented builds on the standards effort being proposed within the OMG, however this effort has not (as of this writing) been issued by the OMG and the involvement of W3C would also help in making a semantic approach to federation viable. This paper explores the possible sponsorship and/or collaboration with W3C on the SIMF effort. More details on SIMF may be found here:

http://www.omgwiki.org/architecture-ecosystem/doku.php?id=semantic_information_modeling_for_federation_rfp

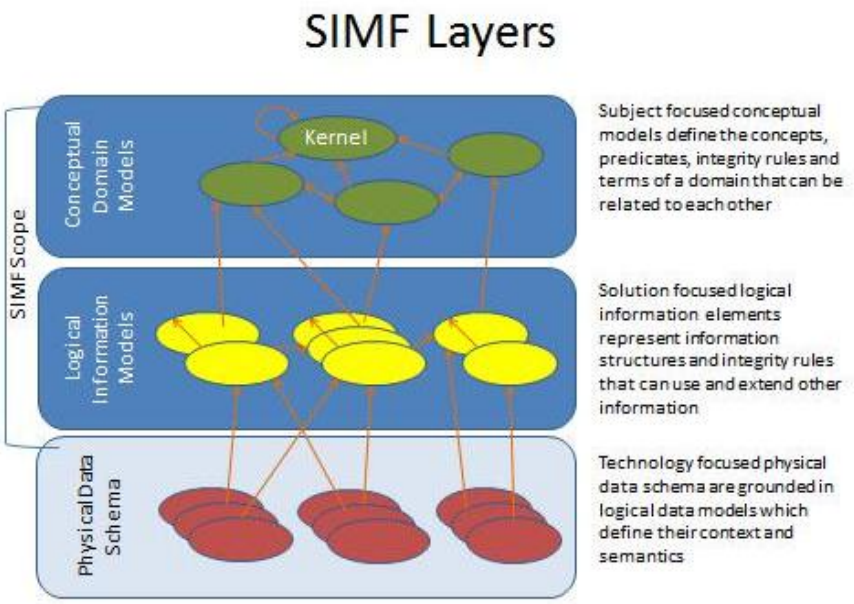
The purpose of SIMF is to help unify and integrate domain information across different authorities, vocabularies and formats. Current conceptual and logical information modeling approaches tend to be focused on a particular information modeling problem, using a particular technical approach. Examples of such technical approaches include object modeling, DBMS modeling, and exchange schema modeling. SIMF seeks to address the problem of information federation by specifying standards for conceptual domain modeling, information modeling and model bridging relationships.

A key assumption of SIMF is that there will be no one single data format, vocabulary, exchange standard or technology for representing information, yet systems using diverse structures and technologies must federate. There will always be diversity such as we see between XSD, RDJ, JSON, SQL, UML, Etc. There will always be different ways to package data for a specific purpose; there will always be different theories on how to abstract real works concepts. The only viable approach to solving the data problem is raising the level of abstraction such that domain concepts, not technology structures, are the cornerstones of interoperability and to provide for bridging between models at all levels. SIMF is intended to provide the framework for this higher level of abstraction as well as its relation to the multiple technologies and data formats that we must federate.

While this draws on semantic technologies the purpose is not the same: SIMF is intended to solve data federation, not the general problems of ontological reasoning.

SIMF Architecture

The diagram to the right illustrates the layers of the semantic information modeling for federation language approach.



Conceptual domain model ***(CDM)***

Conceptual domain modeling primarily addresses the Semantics, concepts and terminology of a domain, e.g., telecommunications, finance or even meta-modeling. It is a model of the terms and concepts of an area of concern (a domain), (not a traditional data model) in that it captures the meaning that usually is not available in a data model while abstracting out data representation and application specific considerations. The objective of the CDM is to capture the semantics of one or more domains as a fully defined set of (potentially federated) concepts, predicates (to express properties about the concepts and to relate them) and integrity rules (constraining instances).

The CDM will satisfy the 100% Principle of ISO TR9007: "All relevant general static and dynamic aspects, i.e. all rules, laws, etc., of the universe of discourse should be described in the conceptual schema. The information system cannot be held responsible for not meeting those described elsewhere, including in particular those in application programs.".

The scope of the CDM includes "controlled vocabularies and "domain ontologies". For a given domain, many conceptual domain models may co-exist, e.g. conceptual domain models that have been developed by different entities and express differing points of view; the MBR part of SIMF makes it possible to explicitly relate the elements such that machine reasoning can be applied.

Semantic relations between elements of these conceptual domain models allow linking those conceptual elements that represent the same concepts, linking related conceptual elements (e.g. supertypes and subtypes). Predicates and integrity rules add to the semantics of each concept.

Conceptual domain modeling can be expressed using graphical or textual notations (potentially in the form of a controlled natural language) (e.g. English, German)..

For example, the CDM may capture the concepts of "mass" and "color" as properties of "physical objects". The same concept of mass may also be known by the term "masse" in German, the CDM will allow for both.

For the purpose of SIMF, conceptual domain modeling is limited to modeling the concepts of a domain. Modeling processes and services is considered out of scope. Future efforts may address other conceptual modeling requirements.

The primary stakeholders of a CDM are the businesspersons, domain experts, business architects, business analysts and ontologists. The businesspersons and domain experts prefer a conceptual domain model

expressed in their familiar jargon and independent of any software technology. The CDM will be optimized for these business and domain stakeholders.

The CDM provides the "*pivot point*" between more purpose-specific layers.

*Logical information models **(LIM)***

Acting as an intermediate between conceptual domain models and physical data schema, the objective of logical information models is to provide a purpose-specific but implementation technology independent view of information in terms of logical data structures. Logical information models are built for achieving certain purposes and provide a way to structure information that is most useful for the intended purpose, application and intended stakeholders.

There can be multiple different ways to represent the same information from different viewpoints and for specific purposes. Each viewpoint may have its own structure, local vocabulary and subset of all possible information in a domain. These purpose specific commitments are made in the LIM.

Elements of a LIM are related to the CDM concepts, predicates and integrity rules they represent and may extend or embed other logical elements.

A LIM model addresses a specific viewpoint and purpose and as such selects those types, properties and relations of interest and structures them for that purpose.

For example color may not be of interest in a particular usage. However, mass measurements may be of interest; a problem-specific LIM may represent a measurement of mass using a particular unit taken at a particular time (noting that both unit and time are also CDM concepts).

Another example, a logical information model applicable to a given company's billing application may address their clients' information e.g. the addresses of their clients and the related purchases, invoices, requests, etc. For its own purpose, the company may have decided to generate from that LIM, both:

- a relational logical information model and a SQL physical data schema for a relational database system and
- another, limited to the client addresses information, hierarchically organized logical information model and a XSD physical data schema as a means to exchange their client information with other companies.

The primary stakeholders of a LIM are software engineers and logical data modelers.

*Physical data schemas **(PDS)***

A physical data schema (or model) describes how to implement a logical information model in the database or exchange format of choice, i.e. defining the application and technology-specific representations of data. There can be many representations of the same logical information model.

For example, an XML PDS may have a composite element type called "ObjectMassMeasurementGrams" representing a mass measurement containing a Real "MeasureMassGrams" (for the mass) and "MeasureDatetime" a Datetime of the measurement. Physical data schemas grounded in logical models expressed and developed using the SIMF language, provide the basis for federation of data defined in those schemas. Such fixed schemas become a particular projection of information for a particular purpose, but not the only way to access the same information. While PDS are out of scope for SIMF, model bridging relations to PDS are part of SIMF.

The primary stakeholders are the software engineer and the physical data modeler.

*Model bridging relations **(MBR)***

The model bridging relations define the *connections between elements of conceptual, logical and physical models* of different layers and within models of a given layer. MBRs also address connections between independently conceived models.

Linking semantics to its different conceptual, logical and physical representations, the model bridging relations are the foundation of federation.

For example, a "shipping weight" in one model may represent exactly the same thing as a "container weight" in another – these can be defined as equivalent expressions of the same thing. More expressive model bridging relations may include the relationship of patterns of concepts to patterns of concepts in another model.

Lessons learned in the SIMF SIG and recommendations

1. The authors recommend using concrete M0 examples to develop the corresponding Conceptual Domain Models at the M1 level, from which the associated M2 level can be derived, via a protocol. Working with this approach saves quite some time in collaborative effort and provides for validation of the M2 model as opposed to postulating it.

2. SIMF satisfies an important need faced by almost every I.T. professional and organization. The authors recommend to apply an open mind to the development of SIMF such that short term commercial interests and religious positions should not influence decisions for the good of a much larger group.

3. The authors welcome a SIMF-like effort in W3C and a possible cooperation between the two groups, with the aim to apply the quality of W3C specifications to SIMF. In addition, the relation of SIMF to W3C efforts at the data representation level will help close caps between logical and physical representations.