Semantic Information Modeling for Federation (SIMF)

Overview V0.2 - 04/21/2011

The Architecture Ecosystem SIG of the Object Management Group (OMG) is in the process of drafting an RFP focused on addressing the information federation and integration problem faced by every enterprise. This is an overview of the SIMF RFP effort.

Purpose

Our ability to share, manage, analyze, communicate and act upon information is at the foundation of the modern enterprise. Information sharing is essential for enterprise supply chains, fighting terrorism and integrating enterprise applications. Yet, this essential capability has remained difficult in information systems which are frequently isolated, stove piped and difficult to integrate. The inability of our systems to share information hampers the ability of our organizations to collaborate and for our processes, services and information resources to work together. Some estimate that more than 1/3 of our information technology budgets are consumed overcoming this “semantic friction” in our systems and that the costs to society from our failure to share and collaborate is many times the systems overhead.

Current standards for information and data modeling are effective at defining a particular data model for a particular application in a particular technology to solve a particular problem. But they suffer when applied to multiple applications for multiple purposes over multiple technologies. They also do not deal with unanticipated needs and opportunities. Most mainstream modeling techniques are challenged when faced with federating independently conceived models, even those in the same modeling tool.

There are several technologies and paradigms directed at information sharing and mapping, yet most of these are ad-hoc or proprietary. There are few standards, and those that exist are focused on efficient access to fixed data sources by relatively static applications. None of the mainstream and standard modeling languages directly target conceptual modeling and data federation effectively. The purpose of SIMF is to put those standards in place.

On the other hand, semantic technologies such as the “semantic web” OWL and common logic provide interesting capabilities for understanding and connecting information but do not do so in a way that resonates with people who define or integrate data models. Semantic technologies frequently expect advanced training in approaches that are not currently mainstream. In addition, the notations (graphical and textual) of semantic technologies are not resonating with “data people”. Standards that provide semantic capabilities in ways that better resonate with users and architects provides an opportunity for a more semantic approach to architecture, one where purpose specific models are linked by common semantics.

Technologies related to SIMF that could benefit by a standard include:

- Master Data Management
- Data Warehouses
- Data Integration and Fusion
• Metadata Management
• Canonical Models
• Linked Open Data
• Controlled Vocabularies
• Ontologies, Semantic Technologies, Semantic Integration
• E/R, XSD and UML Modeling
• Data Mapping and ETL tools
• ISO 11179
• National Information Exchange Network (NIEM)

The issues with data sharing can be roughly subdivided into: infrastructure, format and semantics. All three are required to share information. Infrastructure is the technology used to move data from one place to another – we can do that quite effectively today. Format is the way data is structured and we are somewhat effective at handling multiple data formats, abet via manual and point-point integrations. Semantics is what the data means, and we are not very good at all at understanding how the semantics of data in independent data sources is related. Differences in terminology, viewpoint and purpose make our inflexible data structures hard to integrate.

The theory of SIMF is that we can better relate the semantics and format of data utilizing three basic tools:

• Conceptual domain models that define the terms and concepts of a subject area - the semantics of the domain. This could also be considered a well-defined business vocabulary.

• Logical information models that define the context and structure of data for specific viewpoints and purposes, related to conceptual domain models. A common vehicle for logical information models are E/R models – but even these suffer from an intrusion of the relational technologies they represent and the dependence on rigid schema. UML models, which can be used for the same purpose, are similarly infected by their object oriented technology roots.

• Bridging relationships between different conceptual domain models, logical models and the physical data schema that actually power our information systems. Our assumption is that the physical data models exist in our technologies, i.e. SQL, XSD and OWL. What do not exist are the bridging relationships that can be used within and between models.

**Scope**

SIMF asks for submissions for standards that provide for conceptual domain models, logical information models and bridging relationships in support of information federation. Such models can then power the infrastructures that manage, share, store, transform and
communicate information. The combination of SIMF models with the underlying infrastructure technologies will complete the picture of information sharing.

A basic principle of SIMF is that there is no one “right” conceptualization or representation of information so we need to be able to define our conceptualizations and representations accurately and then relate them, through bridging relations, to other ways to conceptualize and represent the same or related information. It is not our expectation that SIMF can solve this problem 100%, but that a partial solution will substantially reduce the semantic friction between information systems. Even a 50% reduction in semantic friction would result in substantial cost reduction, increased usability of data and correspondingly more effective collaborative processes and services.

The purpose of SIMF is to request submissions for a standard in support of information modeling and federation that will help unify and integrate data across different authorities, vocabularies and formats. The standards will include a model of how information is modeled, or a “meta model” which has very well defined semantics – semantics grounded in formal logic, but formal logic will not be required of those using it. The standard will also include graphical and textual languages that are user-friendly, tuned to the way people think about their information. To exchange information the standard will include a technical exchange format that will support a “linked open data” approach to information models.

Finally, the SIMF standard must itself be integrated – so bridging relations will be required between the SIMF model and UML, E/R, RDF/S and XSD data schema. You must be able to use SIMF with these mainstream technologies.

The intended user of SIMF based tools will be anyone who defines terms and concepts; business architects, data analysts, ontologists, systems architects and data fusion experts. Essentially, anyone who defines or uses structured vocabularies or information. The same resources are also applicable to the extraction of structured information from unstructured resources, such as documents, but unstructured information is out of scope for SIMF.

All ingredients to specify this standard are available today. No fundamental research is required. Even stronger, declarative information transformation has been in practical use since over 25 years in critical business applications using relational technology. What SIMF hopes to do is bring together best practice and best technology into a standard that will make a substantial contribution to solving the data sharing and federation problem.

**SIMF Layers**

SIMF uses a classical layering that is been developed and proven for decades. What SIMF is adding is the explicit language standards for representing the conceptual and logical models as well as the way to make meaningful relations between them. Being able to make relations between elements of the models provides the basis for federation – where concepts are the same or related in different models, semantic relations capture these critical connections. Traversing these connections provides the basis for information federation. Mapping down from conceptual through to logical and physical models provides a basis for defining information once and representing it in many technology forms.
In that many physical data models already exist, SIMF is not inventing more physical data formats but providing a way to make connections from these existing physical schema to the logical layer. SIMF may also be used to produce schema in these existing physical schema languages based on logical information models using the model driven architecture (MDA) patterns.

The diagram above illustrates the layers of semantic information modeling for federation:

- **Conceptual domain models** (CDM) capture the terms and concepts of a subject area – it is a model of the domain or business. The scope of the CDM includes “controlled vocabularies” and “domain ontologies”. Multiple such conceptual domain models exist and there may even by multiples for the same subject area, perhaps independently conceived. Semantic relations between elements of these conceptual models define where they represent the same or related concepts and also help to define the semantics of each concept. Capabilities from semantic technologies will be used to further refine a CDM, but expressed in more user-friendly terms. The CDM is a model of the domain, not of information and is therefore the most general and reusable layer. For example, the CDM may capture the concepts of “mass” and “color” as a 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. Future standards may extend SIMF to include other aspects of the domain model such as process and services.

- **Logical information models** (LIM) define information that captures concepts in a particular way – it is a model of information. There can be multiple different ways to represent the same information from different viewpoints. Elements of a LIM are related to the CDM concepts they represent and may extend or embed other logical elements. The LIM adds structure and context to concepts in the CDM and may select
properties and relations of interest in a specific viewpoint – for example color may not be of interest in a particular usage. However, mass measurements may be of interest; a 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).

- **Physical data schema** (PDS) define technology specific representations of data – there can be many representations of the same logical information element or domain concept. 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 schema grounded in logical models with SIMF provide the basis for federation of data defined in those schema. Such fixed schema become a particular projection of information for a particular purpose, but not the only way to access the same information.

- **Model bridging relations** (MBR) define the connections between models – the orange arrows in the above diagram. Information models may be large and need to be modular such that independent concerns are not conflated. In addition, the reality is that no model will serve all purposes and be fully correct. Independent information modelers will invariably come up with different ways to conceive of, name and represent the world. Model bridging relations provide the connections between elements of different modules – they 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.

With physical schema grounded in conceptual and logical models software systems have a basis for federating information in different data stores. The physical data schema do not have to be changed, they need to be related to the concepts they represent.

**The SIMF Kernel and formal languages**

Concepts are defined in SIMF by using and extending other concepts. So where do fundamental concepts come from? Fundamental concepts may be defined in natural language (preferably controlled natural language) and/or defined in one or more formal languages such as Common Logic, OWL or “Z”. Many domain concepts will stop at well-defined natural language definitions, but some models, and the SIMF kernel model in particular, must have a formal definition.

The SIMF kernel is a set of fundamental concepts that have a formal definition and are used or extended to define other SIMF concepts. Precedent for this has already been set in other OMG standards such as “fUML”, which formally defines an executable subset of UML in common-logic (ISO Common Logic 24707 [ISO/IEC]).

The SIMF kernel is intended to be small but sufficiently expressive that other concepts can be defined precisely and intuitively. The SIMF kernel must be sufficiently expressive to represent its self. Note that users of SIMF may be unaware of this formal grounding yet will benefit from the consistency and precision of the SIMF language.
**SIMF Notations**

Modeling is made real to users in terms of graphical and/or textual notations. SIMF will have at least one graphical and one textual notation for creating CDM’s, LIM’s and MBR’s. These notations will be user-focused, intended to leverage the way people think about information while being clean and precise. The SIMF notations will build on, not abandon, successful visualizations such as E/R and UML diagrams.

Graphical notations are useful for some purposes but textual (controlled natural language) and grid (Spreadsheet) notations are also useful for many purposes. SIMF will include a set of ways to view conceptual and information models that are all based on the same underlying concepts.

**SIMF Use Cases**

The following use cases are intended to provide examples of where SIMF may be employed. [More Needed]

**Cross-organizational data federation for ISE**

Data sharing has been recognized as a key enabler for purposes as diverse as fighting terrorism to financial transactions. One example is the U.S. Information Sharing Environment, or “ISE” ([http://isegov](http://isegov)).

The mission of ISE is:

| The ISE provides analysts, operators, and investigators with integrated and synthesized terrorism, weapons of mass destruction, and homeland security information needed to enhance national security and help keep our people safe. |

The area where SIMF addresses ISE is in enabling integrated and synthesized information. Currently ISE depends on fixed schema for information sharing; NIEM and UCORE. These, in essence, provide common XML schema to which every system must adapt, that adaptation is manual. Of course having standard schema is a big step ahead of no standards at all.

With SIMF, ISE would be able to define a conceptual domain model for national security, derived from the great work that has already been done for NIEM and UCORE. The ISE-CDM could then support the mapping of information found in existing systems to one or more exchange schema like NIEM – it would be far easier to share information without requiring expensive point to point adaption. In addition, with the semantics well defined in the CDM the accuracy of information exchange would be improved and, in some cases, verifiable.

It is accepted that technologies change, if ISE would like to embrace a technology other than XML schema it would be very simple to produce new technology mappings from the ISE-Logical Information Model (LIM). For example, there is already interest in a semantic web (RDF) representation of NIEM.
Integration of data feeds between multiple financial systems
A large financial institution has a critical need to better integrate systems in support of mortgages. Not only has this business area suffered financial distress, there is the need for new federal reporting, new analytics and integration due to acquisitions. Existing systems need to be supported as does a re-engineering of one of the key applications.

The financial institution would like to employ both semantic technologies and model driven architecture in a scalable enterprise solution. They have multiple layers of existing middleware specifications, XML schema for use in web services, event brokers, etc. Most if not all of the existing systems and technologies still need to be supported. There are dozens of enterprise systems involved and hundreds of small applications and spreadsheets.

The approach chosen is to create a “canonical model” of the mortgage domain and then map data into and out of that canonical model using data mapping tools. Unless they “go proprietary” and/or utilize tooling designed for other purposes, there is no standard tooling for the job.

SIMF would be a direct fit for this application, providing a language to create a business vocabulary for mortgages (the CDM) and then create reusable information elements for the transmission and storage of this information in a LIM. SIMF semantic relations would then specify the mapping between their logical information model and the many schema used by existing systems and middleware. These models would be the “front end” to the data messaging and integration middleware.

Providing a federated view of exiting data for analytics
{todo}

Federating process models
The object management group (OMG) has multiple standards related to process: BPMN, UML Activities, BPDM, SysML and SPEM. Each of these was created (within OMG) as independently rooted meta models. There is no direct and standard way to share information between them, there are no concepts shared between them. This causes problems for users who would like to use OMG standards together – a case in point being the “UML Profile for DoDAF and MoDAF” (UPDM) an architectural framework standard sponsored by the U.S. DoD. A stop-gap tactic has been to define a UML profile for BPMN, but it is clear this is not the right approach long term.

The approach enabled by SIMF would be to create a conceptual domain model for process – the domain in this case being process modeling. This would capture the major concepts of process, certainly all that are common between BPMN, UML, BPDM and SPEM. A LIM for process would then capture the common ways these process concepts are organized and contextualized as the bridge between these standards. By bridging the standards we also provide the information necessary to bridge users process models and, ultimately, executing business and technical processes. For example it should be possible for a process defined in BPMN to use information defined in SPEM and call sub-processes defined in UML. SIMF would represent the semantic bridge between processes, process models and process modeling languages.
In this case most of the work to identify process concepts has already been done – what is required is to “lift” those concepts, as we discover and understand them, into a semantic model that is the basis for relating process information. Note that an attempt was made in BPDM to do this in UML but it proved difficult – we didn’t have the right tools for the job.

**Mandatory Requirements**

The following are the requirements for SIMF submissions.

**SIMF Kernel Model**

SIMF shall specify a “kernel” model that includes a set of kernel concepts. The set of kernel concepts shall be those that are sufficient to precisely and intuitively to define the kernel model and all other models in SIMF (the CDM, LIM and MBR). The kernel model shall be defined in terms of one or more existing formal languages such as common logic, “Z” or OWL. All statements made in the SIMF model should have a precise and well defined mapping to the SIMF kernel without information loss.

**Language for Conceptual Domain Modeling (CDM)**

SIMF shall use the SIMF kernel to specify a model for conceptual domain modeling (CDM). This model shall include the set of concepts necessary to define domain concepts in support of the logical model and federation. Proof of generality shall be provided in the form of at least 4 exemplar models taken from different domains. The following features of conceptual modeling shall be supported:

- Types and instances
- Properties
- Associations and hierarchies of associates providing semantic intent
- Constraints
- Expressions
- Composition
- Units and Quantities
- Parameterization of types and patterns
- Multiple names for concepts
- Open and closed world interpretations
- All concepts should be identifiable such that artificial reification of concepts is not required
- Use at any “meta level”, i.e. for data, data definitions or language definitions (M0, M1, M2)

Note that additional detail on these requirements may be found in this white paper: [http://www.omgwiki.org/architecture-ecosystem/doku.php?id=requirements_for_federated_modeling](http://www.omgwiki.org/architecture-ecosystem/doku.php?id=requirements_for_federated_modeling)

**Language for Logical Information Modeling (LIM)**

SIMF shall use the SIMF kernel to specify a model for logical information modeling (LIM). The LIM language shall be capable of representing data context, data structures and viewpoints as may be found in existing representative data descriptions. The LIM language shall not be bound to any particular data representation language but should have sufficient detail and precision to support production and federation of data schema with additional parameterization and bridging relations.
The LIM language should be related to the CDM language such that one or more concepts defined in a CDM can be used to define the semantics of information elements one or more LIM models. The degree of integration and similarity between the LIM and CDM is a design choice of the SIMF submitter.

**Language for Model Bridging Relations (MBR)**

SIMF shall use the SIMF kernel to specify a model for Model Bridging Relations (MBR).

The abstract syntax and semantics of the SIMF language shall include capabilities to semantically relate identical and similar information concepts that have been independently conceived and represented in the same or different information models using the same or different information modeling languages. This shall include differences in name, structure, representation, property sets and underlying semantic theories.

SIMF assumes that models at all levels (CDM, LIM and Schema) may be separated into independent or semi-independent modules and that many of these modules may be independently conceived. The SIMF MBR language will enable these independently conceived models to be federated such that the similarity and difference between elements defined in each can be understood and captured in a model. The set of model bridging relations shall be specified by SIMF submitters, the following are examples of possible bridging relations:

- “same as” – subject elements represent exactly the same concept
- “equivalence” – the subject elements are representations of the same real-world thing or set of things but may have different ways to conceive of those things
- “disjoint” – the real-world things represented by the elements are different, the intersection of their extents is empty
- “projection” – an expression that projects an element or a pattern of elements onto another element or pattern of elements
- “narrower than” – a concept that is a more narrowly defined variant of another
- “broader than” – a concept that is more broadly defined than another
- “subset” – a set that is a subset of another set or type
- “superset” – a set that is a subset of another or type
- “union” – the union of the members of 2 or more other sets or types
- “intersection” – the intersection of the members of 2 or more other sets or types

The degree of integration and similarity between the LIM, CDM and MBR is a design choice of the SIMF submitter.

**SIMF itself is modular.**

The concepts that are defined in the abstract syntax and semantics for the SIMF language shall be defined using reusable language specification modules. Reusability of these modules for other purposes will be a factor in evaluating the quality of the approach but proof of generality will not be required.
Support for Views and Viewpoints.
The SIMF LIM language shall define and provide for a concept of “views” and “viewpoints” as informally described above. The SIMF language shall provide for the same information to be related and synchronized across views that share the same underlying concept. Views and viewpoints shall be able to be dynamically created by architects by assembling the models appropriate to particular stakeholders and stakeholder communities.

A viewpoint specifies a reusable set of criteria for construction and selection of model elements, addressing particular stakeholder concerns. A viewpoint may be partially expressed through the metamodel elements that it exposes, but may also include less structured expressions such as usage patterns and diagramming rules.

A view is a representation of part of a system that conforms to a viewpoint.

Interchange Format.
The SIMF language shall define or utilize a concrete syntax for the exchange of models in its abstract syntax and formal semantics.

Support for information model notation.
The SIMF language shall define at least one diagrammatic and at least one textual concrete syntax that represent the user viewpoint for specifying the SIMF languages – the CDM, CIM and MBR.

The SIMF notations are intended for use by a wide range of users from domain experts to architects and data modelers. Precision, intuitiveness and building on familiar diagrammatic constructs are required.

Use of OMG “Diagram Definition” is the expected form of expression for model notation. If, for any reason, DD is not sufficient submitters may suggest modifications to DD or, in extreme cases, a replacement.

Support Semantic Grounding.
The SIMF CDM language shall support semantic grounding of concepts but shall not require that all concepts are formally grounded. Where there are informal but accepted common concepts the SIMF language shall allow utilization of those informal concepts and definitions. Domain models, languages and viewpoints may have their own “private” concepts that have no grounding at all.

By grounding we mean that each set of statements in SIMF should correspond to one or more statements in the formal language.

MOF Compatibility
Existing MOF models expressed in XMI shall be usable in the SIMF language, either directly or via a transformation.
Web Accessibility
The SIMF language shall provide for modeling concepts and model content to be web addressable resources, have a unique web identity and be de-referenceable based on that identity.

Bridging relations to E/R, XSD, UML and RDF/S
SIMF must itself be federated with common forms of defining information. SIMF will define MBR models that bridge the SIMF model with existing models for E/R, XSD, UML and RDF/S. These bridges are required to capture the common semantics between SIMF and these other standards, not to capture every possible detail of them. In keeping with the SIMF philosophy, it is a federation, not a translation.

Grounded in explicit, concrete examples
The SIMF language shall be explicitly validated by a representative set of examples that demonstrate its applicability to the definition, extension, federation and integration of information models. There shall be a minimum of 4 examples drawn from different domains.