Semantics for Big Data
AAAI Technical Report FS-13-04

Semantic Interoperability in the Oil and Gas Industry: A Challenging Testbed for Semantic Technologies

Wolfgang Mayer, Markus Stumptner, Georg Grossmann, Andreas Jordan
University of South Australia
Advanced Computing Research Centre
Mawson Lakes, Adelaide, Australia
Firstname.Lastname@unisa.edu.au

Abstract

This paper outlines some of the inherent difficulties present in large-scale standards-based semantic interoperability in the Oil and Gas industry. This domain in particularly interesting for semantic interoperability, as the complexity is manifold: data sets are large, span many different domains, are modeled and represented differently in various standards, which have evolved considerably over time. We outline the main challenges with respect to sustained interoperability and advocate that the interoperability scenarios could serve as an interesting test bed for evaluating semantic interoperability techniques.

Semantic Interoperability in Heavy Industries

Interoperability between information systems has been a long standing challenge for software engineers and end users of the resulting applications alike. In particular in heavy industry, such as the Mining and Natural Resources, vast amounts of data related to the construction, operation, and maintenance of plants and other equipment are incurred, as digital sensors can often provide readings at multi-millisecond intervals, and a plant will include thousands of sensors. Moreover, the design, control, and operation of equipment are typically distributed among several companies and vendors, locations, organizations, and personnel, each of which relies on its own business processes, applications, and data storage technologies. However, effective event-driven and auditable processes are required that coordinate the activities across a wide variety of roles, from central management functions such as accounting and planning, to maintenance, control, and supply chain. Achieving this goal demands semantic technologies that enable domain experts to define the information pertaining to their point of view of the system as a whole and synthesize semantics-preserving editable views from a global, detailed information model held in a federation of databases.

The problem of segregation of data held in isolated vertical "silos" within various organizations is particularly pronounced in the Oil and Gas industry, where information related to design, operation, and maintenance of large scale equipment must be reliably exchanged and integrated. Moreover, software and hardware alike are governed by a large number of separate organizations, standards, and vendors. However, to date, little coordination between vendors, owners and operators of plant, equipment and software used for design, operation and coordination of maintenance activities exists.

Standards-based Interoperability

The Open O&M initiative has been established to combat the proliferation of proprietary data formats and software interfaces that have been hindering information exchange and application interoperability in the Oil and Gas industry, with the goal of establishing *semantic interoperability* between information systems. In their working paper (The OpenO&M Initiative 2012), the "big bang" approach to information handover, unstructured and proprietary data formats were identified as three of the main obstacles to data exchange. Through developing best practices and standards for information exchange, the initiative has been attempting to improve the handover of information between various organizational entities. Their efforts focus on methods for the continuous exchange and integration of data in a well-defined standard format.

One of their main goals is integrating the data held in disparate systems such that the meaning of data is preserved and accurately propagated across a federation of various information systems. The diversity and volume of data render this a formidable challenge.

To this end, standards such as ISO15926 (FIATECH 2011) and MIMOSA CCOM (MIMOSA 2010) have been devised that aim to define a common domain ontology, data model, and data formats that can be used to organize information and orchestrate information exchange processes. Both standards are based on well-established technologies: ISO15926 relies on a 4-dimensional (perdurantist) approach to information modeling and STEP/EXPRESS, RDF, OWL, and first order logic for information representation, whereas MIMOSA CCOM employs traditional entity-relationship modeling principles founded in an object-oriented data model expressed in UML. Although both use XML technologies for data exchange, interoperability between the standards remains a contentious issue. The standards are extremely generic in nature, each employing its own layered modeling approach combined with extensible "reference li-

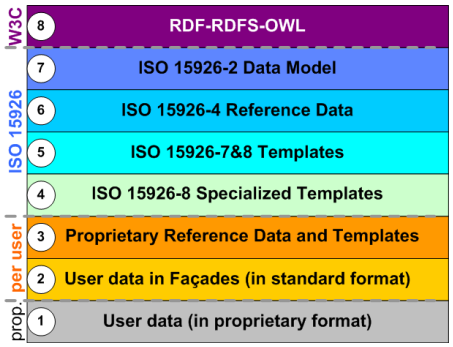

Figure 1: The ISO15926 stack (FIATECH 2011)

braries" that define taxonomies of various domain aspects and their relationships. Figure 1 illustrates the layered approach of ISO15926, where layer 7 provides a generic data model (ca 200 entities) within which an extensive library of "standard" domain specific entities (currently ca 100,000) are defined and continuously extended. These definitions are sourced from various related standards and contributions from participating organizations. In addition to the standard reference library, individual organizations can furnish their own extensions. The actual user data (comprising information related to design, operations, and maintenance of plants) is held in proprietary systems in layer 1. Communication between federated systems is accomplished through standardized "templates" expressing (declarative) chunks of information. Wrappers ("façades") are required to translate between the standard representation and the proprietary applications.

Challenges for Large-Scale Interoperability

Standards alone, however, are not a satisfactory solution, since their generic nature has led to a situation where multiple tool vendors comply with the standard albeit providing mutually incompatible implementations that are unable to inter-operate. Moreover, the precise meaning of elements in the standard is often difficult to ascertain (Smith 2006), and identity management of information entities across system boundaries remains an issue as systems may not all rely on the same identity criteria and keys for identification.

Methods and tools that can establish continued, reliable, and bidirectional information integration within a federated architecture of information systems remain to be desired. Current platforms, such as iRing[1], are limited to particular implementation technologies in the end user's system and are not sufficient to seamlessly integrate other systems; in particular since the creation of wrappers for proprietary systems is difficult due to often considerable mismatches between the internal and external representations of information, evolution of the standards, and ambiguity in representation. The current tool landscape for creating wrappers and mappings between internal and external information representation are relatively primitive and do not provide satisfactory support for the discovery of mappings and transformations between complex data structures at a large scale. Moreover, systematic mechanisms to versioning of transformations and mappings remain to be desired.

In the following we briefly outline some of the challenges for large-scale standards-based interoperability pertaining to (but not limited to) the interoperability of MIMOSA and ISO15926.

Reference model alignment. In an attempt to align the reference data libraries of MIMOSA and ISO15926, we employed schema matching tools, in particular COMA++ (Aumueller et al. 2005), to the taxonomies comprised in the reference libraries. Unfortunately, even when selecting the computationally less expensive lexical matching techniques provided by COMA++, we were unable to find suitable alignments given reasonable resource requirements. Only after considerable effort of manually analyzing and partitioning the reference libraries into specific subsets were we able to align some entities. We conducted a smaller study on the subset of concepts related to Heat Exchangers. However, the results were discouraging due to the fact that useful candidate matches could be found for only 4% of concepts. For the remaining entities there were no candidate matches or multiple matches with similar confidence (maximum confidence was 56%). Enabling structure-based matching algorithms to improve the outcome was unsuccessful due to scalability issues and considerable differences between the structural organization of the input taxonomies. Given that lexical similarity was insufficient, and the fact that the taxonomies merely express specialization of various concepts but do not furnish a comprehensive ontology that includes attributes and other relationships that could be exploited in matching, we do not anticipate elaborate ontology matching approaches (Euzenat et al. 2011) to achieve a substantially different outcome.

Inferring candidate mappings from instance data is also fraught with problems arising from vast differences in representation formats of different standards and the absence of lexical similarity of entities. Moreover, the scope and level of detail varies between the standards and no comprehensive shared data model or data structure representation exists across applications. If one accepts the premise that there is a shared understanding of reference models and data representations among the stakeholders of information systems, approaches like Data Tamer (Stonebraker et al. 2013) could serve to distribute some of feedback workload among stakeholders, provided that the matching mechanisms can be extended to cope with heterogeneous complex data structures and multiple levels of abstraction.

To the best of our knowledge, no schema or ontology matching tool exists that would be able to —without relying on excessive user feedback and technical knowledge— simultaneously identify unique matches with high confidence in (i) heterogeneous data models (ii) that are represented using different structure and encodings with (iii) disparate reference libraries where (iv) little or no lexical similarity is present (v) in a scalable manner.

[1] https://code.google.com/p/iring-tools/

Multiple Viewpoints. Aligning representations stemming from multiple domains is further complicated by the fact that there is no single shared conceptual model among applications and stakeholders. Standards such as ISO15926, MIMOSA CCOM, and IBM Reference Semantic Model[2] all provide different yet overlapping viewpoints of the same domain information, albeit at various level of abstraction. However, comprehensive interoperability requires to span all levels of abstraction, and restricting interoperability to the common denominator of several standards will not be sufficient to satisfy all stakeholders' needs. Therefore, it is necessary to semantically align the various viewpoints such that data at one level of abstraction can be linked to other levels. Contemporary approaches to ontology alignment and schema transformation however are focused on relatively simple structural transformation and have difficulties bridging shifts in viewpoint without loss of information. Automatic methods for obtaining suitable linking and transformation mechanisms from a given set of standards/ontologies remain to be desired.

Language Heterogeneity. The formal languages employed in various standards is diverse, not just among standards, but also within. For example, CCOM relies on UML, XML and XSchema to express data formats, whereas ISO15926 uses a combination of STEP/EXPRESS, XML, RDF, OWL, and predicate logic. Matching and aligning ontologies and data models authored in multiple languages is not adequately addressed by current toolkits.

The differences in formal semantics and expressivity of languages adds further to the complexity of the problem. For example, while OWL is sufficiently expressive to define various taxonomies, certain invariants in, for example geometric models, cannot be expressed directly (Motik, Grau, and Sattler 2008). To capture such model properties, comprehensive axiomatizations, for example in the style of PSL (Bock and Grueninger 2005), are needed. Authoring and validating such precise specifications is not only difficult (Beeson, Halcomb, and Mayer 2011), it is furthermore non-trivial to preserve such invariants in the translation to other models.

Evolution and incremental change. The evolution of standards over time constitutes an interesting challenge. Both ISO15926 and MIMOSA CCOM have evolved considerably over time. The former has shifted in representation from STEP/EXPRESS to RDF, OWL and FOL, and has recently begun to include complementary modeling principles adopted by Gellish (van Renssen 2005). MIMOSA has also undergone considerable changes, where the initial purely relational data model has been abandoned in favor of an object-oriented meta-model approach complemented using a reference library. Since both standards keep evolving, it is essential that any alignment relationships and transformation be carried forward as much as possible as either standard is updated incrementally. Model driven approaches to change propagation and devising and maintaining the suitable inter-standard links could serve as the basis for such an approach. However, most contemporary approaches, including ours (Berger et al. 2010), either require users that are experts in model transformation, or cannot cope with complex transformations or incremental evolution.

Consistency and Quality. Since the data models underlying the ISO15926 and MIMOSA standards are authored jointly by a large number of individuals, and the size and complexity of standards prevents most contributors from comprehending their entirety, quality issues may arise. For example, inappropriate specialization or classification relationships may be introduced or relationships may be forgotten, duplication and inconsistency may be introduced, and in appropriate level of details may be provided for some parts of the reference model. While methods and tools for assessing data quality in purely relational schemas is well understood, similar techniques for complex ontologies and theories are still in their infancy and topic of active investigations[3]. In order to ensure the standards remain consistent and of high quality, appropriate measures, techniques and tools should be devised that support knowledge engineers and domain experts at all levels of expertise in authoring and validating modifications to the standard.

Moreover, incremental quality assessment tools and instance migration mechanisms are desired to cope with situations where the best practices for modeling change, as can be expected by the on-going introduction of Gellish into ISO15926 and the adoption of the ISO15926 reference model within the MIMOSA consortium.

Bidirectional Transformations. Aligned meta-models and reference models of standards are only part of the solution and must be complemented with solid execution frameworks supporting the execution and orchestration of data exchange processes in a federated architecture. Different from typical semantic integration scenarios one would find in business intelligence or ontology alignment, semantic interoperability in the Oil and Gas domain requires bidirectional data exchange mechanisms, where information is not only obtained by querying data sources, but also update data in the distributed architecture. Since wrappers are the predominant method of interfacing with legacy systems, wrappers must support both read and update transactions. Here, key issues are the specification of suitable bidirectional transformations that maintain both the identity of the affected data element and the intended meaning of the updates. For example, idempotent updates should not result in duplicate information being introduced in a legacy system when applied more than once. While identity management is less of an issue in domains where identity of concepts and instances is either defined by unique identifiers (such as keys and URLs) or by similarity of attribute values, the distributed domain and varying data models present challenges

[2] http://pic.dhe.ibm.com/infocenter/iicdoc/v1r4m0/topic/com.ibm.iic.doc/model_man/rsm.html

[3] The Ontology Summit 2013, a three month open annual event jointly organized by Ontolog, NIST, NCOR, NCBO, IAOA and NCO NITRD, was dedicated to the topic Ontology Evaluation Across the Ontology Lifecycle. http://ontolog.cim3.net/cgi-bin/wiki.pl?OntologySummit2013

yet to be resolved. The fundamental differences inherent to 4-D data models used by the ISO15926 standard and the entity-relationship model employed in MIMOSA further adds complexity. Currently, it is not clear how comprehensive identity and key management can be achieved without relying on costly explicit key mapping tables in each wrapper component.

Validation and Verification. Moreover, the verification and validation of transactions and mapping is difficult in a distributed context, where both the transformation and the underlying standards are subject to evolution. Due to differences in data format and scope of systems it is difficult to establish a common testbed to be used for compliance checking and regression testing. Moreover, testing alone does not guarantee that the transformations operate correctly for unseen input. Complex transformations embedded in wrappers may hence be difficult to validate. Although formal verification mechanisms could provide such assurances, these generally require experts and often suffer from poor scalability. Finally, there are no unanimously accepted consistency criteria for semantic aspects beyond syntactic conformance and the predefined entity identifiers from the reference model.

Scalability. The scalability of existing systems is problematic due to the overly generic meta models of standards like ISO15926, where multiple levels of concept instantiation may exist and virtually all relations are reified. Further work is required to create and maintain efficient implementations that can automatically cope with schema changes, extensions to the standards' ontologies, and migration of instances in legacy systems.

Semantic Standards Interoperability

Our recent contributions to standards-based interoperability lie in a multi-layered model-driven framework for specification and execution of model transformations (Berger et al. 2010; Jordan et al. 2012). We employ semantic technologies within an MDA-based framework in which multiple levels of linked specifications capture relationships between elements in the various standards and their meta-models. As such, the layered architecture enables us to leverage equivalences on the meta-model level which induce the set of possible and valid transformations on the lower levels. This model-driven approach allows us to concisely specify domain-specific transformations that cover both structural and semantic transformation, including aggregation, differences in coding and other arithmetic transformations. We found that this flexibility is indispensable to overcome comprehensive differences in structural representation and encoding in the different standards.

Our software suite includes a domain-specific editor for defining mappings, and a service-oriented architecture implementation that realizes the execution engine for mapping between the ISO15926 and MIMOSA standards as well as other formats and APIs, including relational databases, CSV, and SAP Netweaver API.

The feasibility of our approach was recently demonstrated in the course of a demonstration conducted jointly with our partners at the ISA Automation conference in Orlando, 2012. A description of the use cases and outcomes is available at http://iringug.org/wiki/index.php?title=GS_OGIDemo_001.

Conclusion

This paper outlines some of the inherent difficulties present in large-scale standards-based semantic interoperability in the Oil and Gas industry and poses interesting challenges for future research. The main issues outlined here concern alignment of large ontologies, distributed authorship and maintenance of reference models, and specification and execution of bi-directional distributed information integration. We have made considerable progress using semantic technologies and model-driven principles, yet we believe that the remaining research challenges could serve as an interesting test bed for evaluating future approaches.

References

Aumueller, D.; Do, H. H.; Massmann, S.; and Rahm, E. 2005. Schema and ontology matching with COMA++. In *Proc. SIGMOD Conference*, 906–908.

Beeson, M.; Halcomb, J.; and Mayer, W. 2011. Inconsistencies in the process specification language (psl). In *Workshop on Automated Theory Engineering (ATE 2011)*, 9–19.

Berger, S.; Grossmann, G.; Stumptner, M.; and Schrefl, M. 2010. Metamodel-based information integration at industrial scale. In *MoDELS (2), LNCS 6395*, 153–167. Springer.

Bock, C., and Grueninger, M. 2005. PSL: A semantic domain for flow models. *Software Systems Modeling* 209–231.

Euzenat, J.; Meilicke, C.; Stuckenschmidt, H.; Shvaiko, P.; and dos Santos, C. T. 2011. Ontology alignment evaluation initiative: Six years of experience. *Journal on Data Semantics XV* 158–192.

FIATECH. 2011. *An Introduction to ISO 15926*.

Jordan, A.; Grossmann, G.; Mayer, W.; Selway, M.; and Stumptner, M. 2012. On the application of software modelling principles on iso 15926. In *MOTPW, Workshops and Symposia at MODELS 2012*.

MIMOSA. 2010. *Open Systems Architecture for Enterprise Application Integration – Version 3.2.2 Specification*.

Motik, B.; Grau, B. C.; and Sattler, U. 2008. Structured objects in owl: representation and reasoning. In *WWW*, 555–564. ACM.

Smith, B. 2006. Against idiosyncrasy in ontology development. In *Proceedings Formal Ontology in Information Systems (FOIS'06)*, 15–26. IOS Press.

Stonebraker, M.; Bruckner, D.; Ilyas, I. F.; Beskales, G.; Cherniack, M.; Zdonik, S. B.; Pagan, A.; and Xu, S. 2013. Data curation at scale: The data tamer system. In *CIDR*.

The OpenO&M Initiative. 2012. OpenO&M Industrial Capital Project Plant Information Handover Best Practices Guide. Technical Report CP-BP001-1.00.

van Renssen, A. 2005. *Gellish: A Generic Extensible Ontological Language*. Ph.D. Dissertation, TU Delft.