A pattern-based approach to ontology-supported semantic data integration

Andreas A. Jordan **Wolfgang Mayer** **Georg Grossmann**
Markus Stumptner

Advanced Computing Research Centre
University of South Australia,
Mawson Lakes Boulevard, Mawson Lakes, South Australia 5095,
Email: joraa001@mymail.unisa.edu.au
Email: (wolfgang.mayer|georg.grossmann)@unisa.edu.au, mst@cs.unisa.edu.au

Abstract

Interoperability between information systems is rapidly becoming an increasingly difficult issue, in particular where data is to be exchanged with external entities. Although standard naming conventions and forms of representation have been established, significant discrepancies remain between different implementations of the same standards and between different standards within the same domain. Furthermore, current approaches for bridging such heterogeneities are incomplete or do not scale well to real-world problems.

We present a pragmatic approach to interoperability, where ontologies are leveraged to annotate prototypical information fragments in order to facilitate automated transformation between fragments in different standards and representations. We discuss our approach based on an example drawn from the oil and gas industry.

1 Introduction

A growing number of organisations are interested in sharing information between their internal information systems and externally with partners or the public sector. A core issue facing these organisations is the information management systems they employ have been built without supporting a common standard for the underlying data models used for storing information and their interfaces. Consequently, creating custom-built software that translates between different interfaces and data models is a predominantly manual task, and current solutions for bridging heterogeneities between information systems do not scale well to real-world problems.

Although standardisation of domain models, interfaces, and information representation can help to reduce this burden, standardisation alone has been insufficient to overcome the interoperability challenge. This is predominantly due to the fact that the standards themselves are heterogeneous and sometimes suffer from ambiguity. In the Engineering domain alone, there exist several competing standards and de-facto standards set by large tool vendors. For example, in the oil and gas industry, the MIMOSA[1], RSM[2], ISO 15926[3] and Gellish[4] standards are all used to represent information about engineering assets and associated maintenance information. However, there exists considerable variation in the scope and level of detail present in each standard, and where the information content covered by different standards overlaps, their vocabulary and representation used vary considerably. Furthermore, standards like ISO 15926 are large yet offer little guidance on how to represent particular asset information in its generic data structures. Heterogeneities arising from continued evolution of standards have further exacerbated the problem. As a consequence, tool vendors formally support the standard, yet their implementations do not interoperate due to different use of the same standard or varying assumptions about the required and optional information held in each system. Not all standards are supported by formal ontologies that would allow one to overcome these heterogeneities automatically. Where formal models are present, their focus is typically on describing individual entities and relationships, whereas dealing with issues arising from their use in larger representation structures, and scalability problems stemming from application of inference mechanisms, are often left aside.

The aim of our work is to build a bridge that ties together the heterogeneous standards and provides automated translation between their information models and concrete implementations. Our hypothesis is that this goal can be achieved only if the information representation fragments used by different tools and standards are captured such that possible translations between equivalent fragments can be inferred automatically. To resolve the interoperability problem not only with respect to terminological heterogeneities but also with respect to heterogeneous usage patterns by different organisations, existing standards must be complemented with a formal model of the elementary "information fragments" and possible use of these fragments within each standard and organisation. Whereas approaches for translating a semi-formal model like Gellish into natural language exist, we aim to extend these ideas to formal complex structures. This requires a formal model of the building blocks, henceforth "information fragments", in the source and target representation and mechanisms to translate between "similar" fragments in different standards.

For this purpose, we intend to construct an ontology that reconciles the concepts and relationships in each standard into a common model and formally express the information fragments found in each stan-

[1] http://www.mimosa.org/

[3] http://15926.org/home/tiki-index.php

[4] http://www.gellish.net/

dard in terms of the common model. Our focus will be on representing information fragments and their concrete representation in different standards such that equivalent representations in different standards can be constructed automatically.

We seek to investigate formal models and inference procedures to determine which are adequate for capturing the intent and composition of concrete data structures on each end of the translation process while reusing as much as possible existing ontologies capturing domain-specific entities and their relations. Whereas (domain-specific) ontologies provided by several standards exhaustively capture domain concepts and relations, we intend to develop mechanisms for reasoning about translations between different representations.

The resulting ontology will help to build adequate translation systems by providing developers of integration tools a catalogue of possible translations between different systems, either at run-time or at design time, when translations for information fragments in the source representation must be formed. Our ontology will also be able to verify completeness of a given translation by identifying fragments and constraints in the source model that cannot be translated precisely into the target model, and fragments that have ambiguous or multiple conflicting translations.

Our research will be predominantly driven by a case study within the oil and gas industry, where information models held in MIMOSA, ISO 15926, and RSM are to be synchronised among a number of systems.

In this paper we make the following contributions:

- we outline our overall approach to automated fragment-based information integration,
- present our approach to linking existing taxonomies and ontologies to our model,
- show how to capture information fragments and associated translation operations in an ontology, and
- show how the resulting ontology can be applied to infer automated translation between source and target representations.

The paper is organised as follows. We present our overall approach in Section 2 and introduce a running example in Section 3. We discuss related work in Section 4 before outlining future work and concluding the paper in Section 5.

2 Approach

Our approach to overcoming heterogeneities in different standards is based on the construction of a unifying ontology that comprises the information fragments used in source and target standards. The ingredients to our approach are depicted in Figure 1: information fragments and transformations are represented in a unifying central ontology, and fragments are linked to concrete representations, so-called "templates", in their source and target representations.

Fragments and Templates. We will use the term "template" to refer to a particular parametric representation of a chunk of information in the source and target representation (which may or may not be supported by an ontology), and will use the term "fragment" to refer to its parametric representation in our

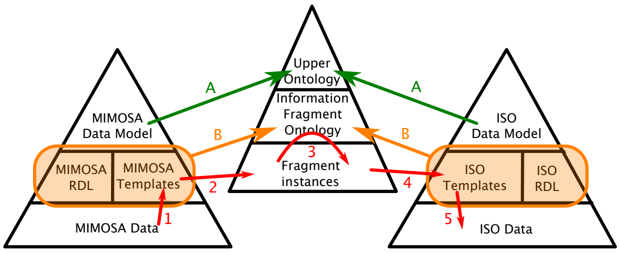

Figure 1: Schematic of the overall translation process on the example of translation between the MIMOSA and ISO 15926 standards.

unifying ontology. Each fragment can be seen as a parametric statement that, when instantiated with suitable values and objects for its parameters, expresses a certain meaning in the concrete representation corresponding ontology. By combining fragment instances representing template instances in the target representation, suitable translation for a given set of template instances in the source representation can be devised.

Ontology of Information Fragments. Central to our approach is the unifying ontology of information fragments and translation operators depicted in the centre of figure 1. It extends a suitable upper ontology in order to foster reuse and to ensure its overall organisation adheres to well-established knowledge representation principles. Each fragment in our ontology is linked with its corresponding template(s) in one or more source or target ontologies. Note that not every fragment will have corresponding templates in each data source. Therefore, it is necessary to find combinations of fragments that are available in the target ontology that correspond to fragments found in the information source. Furthermore, it may be necessary to query information from the target in order to obtain information (such as object identifiers of existing objects) while instantiating target fragments.

We intend to construct an ontology that captures the properties and constraints of fragments in a formal manner such that equivalent instances of fragments can be identified and synthesised. In order to reason about transformations, it is necessary to capture the constraints that govern a template's use in the target ontology and the constraints on its parameters. In addition, a library of conversion operators that transform between representations of basic data types will be included in order to enable automated construction of equivalent fragment instances where data types differ. Our intent is not to reinvent yet another ontology of entities and relationships, but to reuse as much as possible existing ontologies and their formal underpinnings and focus on the representation of information fragments, their parameters and constraints.

Naturally, effective fragment representation and translation depends on a unified model of the underlying entities and relationships. We intend to obtain such a model by extending a suitable upper ontology with concepts and relationships relevant to our domain. These will be drawn from the data models underpinning each standard (arrows labelled A in Figure 1), along with a unified taxonomy of entities/relationships defined in each standard (arrows labelled B).

We intend to leverage as much as possible existing large scale ontologies, especially those that have been developed by the oil and gas industries, but the latter cannot cover all bases. While the use of the ISO15926 ontology has been suggested as a universal upper ontology, it has been shown to suffer from significant intrinsic shortcomings Smith (2006), as well as being based on a formalism that has been shown to make it difficult to concisely express engineering knowledge Felfernig et al. (2003). To express standard engineering data sheets, ISO15926 therefore relies on a first order logic platform different from its core ontology specified in Part 2 of the standard, formally specified but not implemented, and subject to creating classic "impedance mismatch" situations for modellers as different formalisms have to be joined in one ISO15926 application.

Whereas we anticipate that lexical matching and specialisation relationships will allow us to largely identify corresponding entities in this domain, we expect that more involved procedures and manual intervention will be necessary in order to address discrepancies arising from different standards and different granularity of representation. Our work will thus highlight where established standards agree and where further work is needed to reconcile modelling heterogeneities.

Translation Process. The overall approach is summarised as follows, as shown by the red path in Figure 1: First, the information in the source repository (MIMOSA) is read and templates are instantiated accordingly (Step 1). Each template is then translated into its counterpart fragment in our unifying ontology (Step 2), where equivalent fragments in the target representation are subsequently identified and synthesised (Step 3). Once a representation in terms of the target fragments has been obtained, each fragment is translated into its corresponding template in the target representation (Step 4), which is then emitted into the target system (Step 5).

One may argue that suitable templates/fragments and their translations should not be defined explicitly but instead be inferred directly from the ontologies underlying the source and target systems. Unfortunately, such ontologies do not always exist explicitly, cover only the information part but not its concrete representation, or are not sufficiently complete to facilitate all needed inferences. Therefore, our approach relies on a pragmatic template-based approach, which also helps when dealing with issues such as tools that do not fully conform or abuse the official standards. Furthermore, current formalisations are not typically rich enough to overcome significant differences in representation, such as property-based vs. relationship-based models, in particular if source and target use different levels of granularity for certain aspects of an entity.

Validation. The ontology and transformation operators can be leveraged to verify the consistency and completeness of translation. Formal analysis of the possible translation sequences can establish if there are multiple possible representations in the target representation, or if some parts of the source representation could not be translated. In both cases, the ontology may need to be augmented in order to eliminate unwanted translations and improve coverage.

3 Case Study

This case study is based on the MIMOSA and the ISO 15926 standards used predominantly within the oil and gas industry for representing and managing the life-cycle data of assets. Whereas ISO 15926 is predominately used for the exchange of design information, MIMOSA is mainly used in the operations and maintenance area. Both standards provide a data model that specifies a generic structure to represent asset information, along with a "Reference Data Library" where the available asset types and their properties are defined.

We consider the frequent use case of hand-over of design information to operations and maintenance that occurs when, for example, an asset is installed on site. Although both standards are designed to overcome the interoperability problem, currently all translation between these two standards is performed manually.

In the following we will briefly illustrate the differences in representation between the two standards, and show that a common framework of information fragments and transformation operations can overcome these discrepancies.

For purposes of illustration, we assume that translation between the MIMOSA CCOM 3.3[5] and the ISO 15926 (ISO 2003) data models is to be performed. However, our approach will be applicable in scenarios where data represented in multiple standards, such as RSM and XMpLant[6], must be utilised jointly to perform a complete translation. For example, the current ISO 15926 standard must be complemented with additional geometry information stored in XMpLant data structures in order to synthesise a complete representation to be consumed by an RSM-compliant tool.

3.1 MIMOSA

The data model of MIMOSA has been defined using the language of the Unified Modelling Language (UML). The various entities in the data model are represented as classes and the permitted relationships between the various entities as associations between classes. The data model is generic in that it allows for extensions to the model by means of a reference library comprising the asset types, units of measure, and other domain-specific entities that may be assigned to properties of classes in the generic data model. The resulting model is serialised into an XML document and can be transmitted via XML Web Service interfaces defined in the standard.

Figure 2 depicts a partial model of a property attached to an asset. The labels 'A' through 'E' indicate instantiated templates used to represent the property *'Ambient operating temperature range'* for the attached asset, i.e. a pump. The templates labelled 'A' and 'B' associate an Attribute with a type. Templates labelled 'C' and 'D' associate the attribute to numerical values whereas the templates labelled 'E' are needed to constrain the magnitude of each attribute type to 'Degrees Celsius'.

3.2 ISO 15926

This standard evolved from a smaller project beginning in 1991 in an effort to provide a generic data model for representing assets throughout their life cycle. One of the goals of this standard is "to facilitate integration of data to support the life-cycle activities and processes of process plants." (ISO 2004). For this purpose, it relies on a generic data model

[5] http://www.mimosa.org/?q=resources/specs

[6] http://www.noumenon.org.uk/category.php?id=NA==

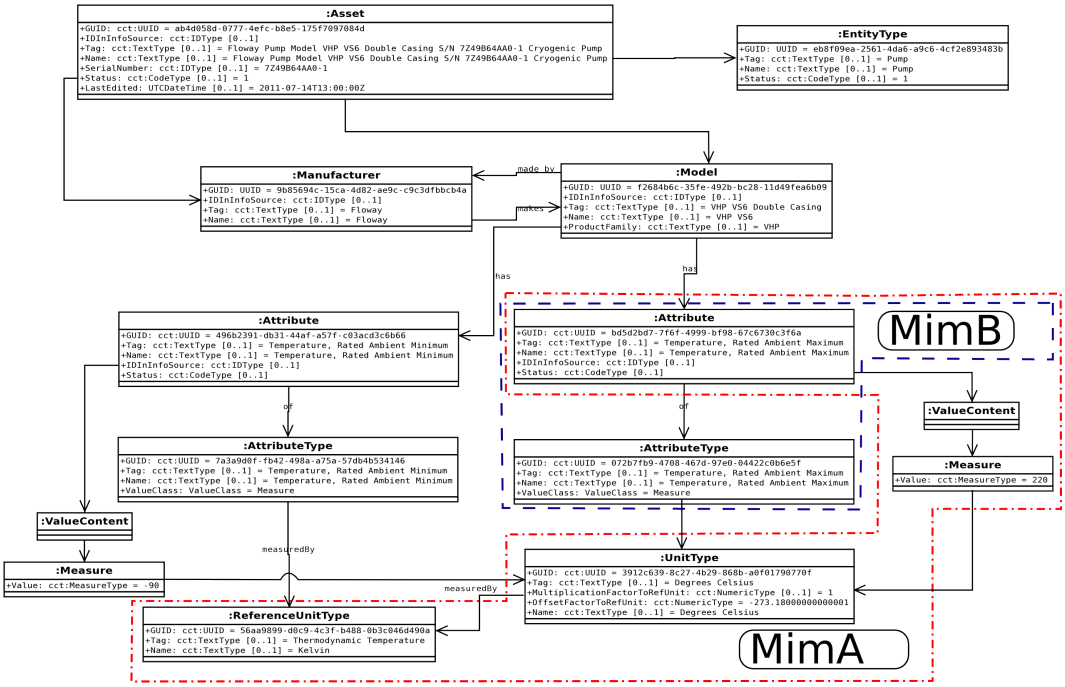

Figure 2: MIMOSA CCOM: Ambient Temperature Range

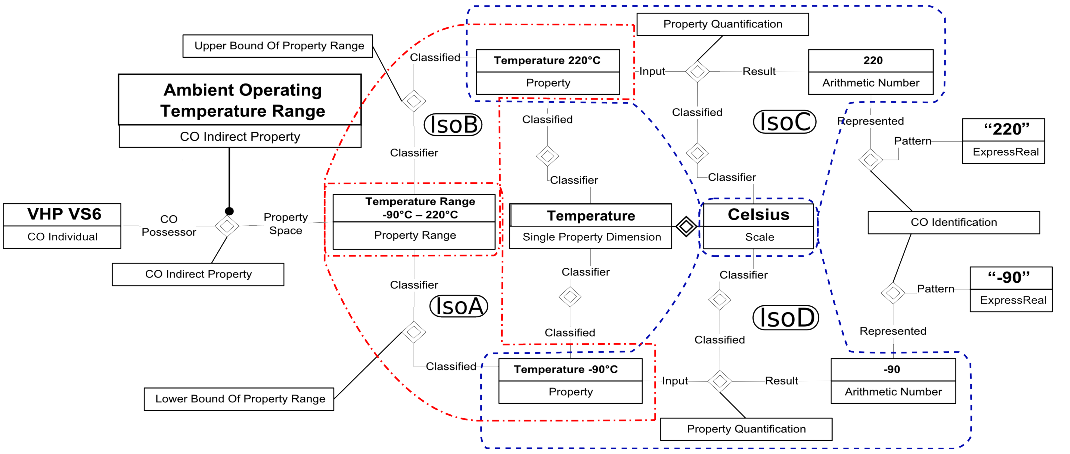

Figure 3: ISO 15926: Ambient Temperature Range

where all properties and facts are expressed as relationships between objects. A catalogue of well-known object classes and relationship classes is defined in its so-called "Reference Data Library", an extensible catalogue accessible via Web Service interfaces. Similar to the MIMOSA library of component types, the ISO reference library associates unique identifiers with each well-known entity in the reference library, thus defining a public "vocabulary" that shall facilitate interoperability. In addition, the standard specifies well-known "templates" that express elementary statements within the object model, and defines a data exchange format in terms of XML and the W3C's RDF framework.

Figure 3 shows a partial model of the same asset property expressed in the ISO 15926 data model. It can be seen that numerous templates are needed in order to state that the property is an interval expressing a temperature in the Celsius scale[7]. Minimum and maximum bounds of the interval are denoted by a classification relationship (*LowerBoundOfPropertyRangeTemplate* labelled 'IsoA' and *UpperBoundOfPropertyRangeTemplate* labelled 'IsoB' respectively), and each boundary property is linked to its value and scale via a ternary relationship (*LowerUpperMagnitudeOfPropertyRange* labelled 'IsoC'). The entire model is composed from templates, where shared parameter assignments manifest as overlapping regions in the figure.

3.3 Transformation

Comparing the MIMOSA and the ISO 15926 representation, it is easy to see that both models follow fundamentally different approaches. Automated translation between the two would be challenging at best, as many elements, such as *PropertyQuantification* and *ClassOfIdentification*, do not have corresponding counterparts in the other model. Furthermore, the libraries providing reference data in both standards differ in many aspects. In our example,

[7] Note for clarity not all templates identified are highlighted

the modelling of temperature scales differs, where MIMOSA follows an indirect approach using a base type and a reference type.

However, these differences can be hidden by translating each of the templates into the unified information fragment ontology. In this intermediate representation, only the intent of the template and the meaning of its parameters are captured, while representation-specific details are omitted.

We illustrate the approach based on a transformation from MIMOSA to ISO 15926. Figure 2 shows a partial model of a temperature range attribute attached to an asset type. The outlined elements indicate two template instances, *MimA* and *MimB*. The former represents the statement that the attribute's value is 220 with unit *Degrees Celsius*, and the latter expresses that the attribute's type is *Temperature, Rated Ambient Maximum*. Both, scale unit and type stem from MIMOSA's reference data library which is assumed to have been mapped into our unified ontology.

Assume that the ontology provides the following statements about properties and their types:

$$MagnitudeOfProperty(x1, x2, x3) \rightarrow$$
$$Property(x1) \wedge ArithmeticNumber(x2) \wedge Scale(x3)$$
$$UpperBoundOfPropertyRange(x, y) \rightarrow$$
$$Range(x) \wedge Property(y)$$
$$PropertyType(x, y) \rightarrow Property(x) \wedge Type(t)$$

Furthermore, assume that the reference data libraries in both standards have been aligned and reconciled into a unified ontology as well. For brevity, we only show the subset necessary to illustrate our example:

$$Type(TRAMax)$$
$$SubTypeOf(TRAMax, Temperature)$$
$$PropertyType(x, TRAMax) \rightarrow$$
$$\exists r : UpperBoundOfPropertyRange(r, x)$$
$$PropertyType(x, t) \wedge SubTypeOf(t, u) \rightarrow$$
$$PropertyType(x, u)$$

The MIMOSA information fragments corresponding to the templates can then be characterised as follows:

$$MimosaType(TRAMax)$$
$$MimB(x, y) \leftrightarrow PropertyType(x, y), MimosaType(y)$$
$$MimA(x, y, z) \leftrightarrow MagnitudeOfProperty(x, y, z) \wedge$$
$$MimosaType(z)$$

Term *TRAMax* represents the mapped attribute type *Temperature, Rated Ambient Maximum.* We use predicate *MimosaType* to indicate that a particular element is mapped from the MIMOSA reference library. This will be needed when translating from ISO15926 to MIMOSA.

Similarly, let the ISO15926 fragments corresponding to the templates shown in Figure 3 be represented as follows:

$$IsoType(Temperature)$$
$$IsoType(SC)$$
$$IsoD(x, y, z) \leftrightarrow MagnitudeOfProperty(x, y, z) \wedge$$
$$IsoType(z)$$
$$IsoA(x, y) \leftrightarrow UpperBoundOfPropertyRange(x, y)$$
$$IsoPropType(x, y) \leftrightarrow PropertyType(x, y) \wedge IsoType(y)$$

Similar to the previous statements, we annotate mapped reference library entities with a predicate that indicates their origin. This will be needed to ensure that only entities are emitted that exist in the target standard. We write *SC* to denote the scale *Degrees Celsius*.

From analysis of the source MIMOSA data, templates have been recognised and translated into corresponding fragments, yielding the following facts in the unified ontology:

$$MimB(Ax, TRAMax)$$
$$MimB(An, TRAMin)$$
$$MimA(Ax, 220, SC)$$
$$MimA(An, -90, SC)$$

We use Ax and An to refer the properties corresponding to the two attributes present in the example model. *TRAMin* represents the type *Temperature, Rated Ambient Minimum* in the unified reference library ontology.

Given these fragments corresponding to the information fragments derived from MIMOSA template instances, we can now infer the equivalent fragments corresponding to ISO15926 template instances that are to be emitted in the target model: From $MimB(Ax, TRAMax)$ we obtain $PropertyType(Ax, TRAMax)$ and, by the subtype axiom, $PropertyType(Ax, Temperature)$. Since *Temperature* is a type available in the target standard, this yields fragment *IsoPropType(Ax,Temperature)*, which can be instantiated as template directly in the target model. Similarly, $MimA(Ax, 220, SC)$ yields $IsoD(Ax, 220, SC)$. Furthermore, $PropertyType(Ax, TRAMax)$ entails that $(\exists r)UpperBoundOfPropertyRange(r, Ax)$, which allows us to instantiate an ISO15926 template associated with $IsoA(r_0, Ax)$. Note that the existential quantification forces us to select a suitable r_0 representing an interval property in the target model. For "green fields" transformations, where the target model is empty, a suitable entity can be generated. Otherwise, the target endpoint must be queried to obtain the needed entity.

4 Related Work

Gellish and the ISO 15926 standard both have constructed comprehensive ontologies and data models for representing engineering assets, processes, and associated meta-data. Common to both is the assumption that all data not represented in other forms is converted into the standard's ontology. In the case of ISO 15926, simple tools have been developed that allow one to create simple mappings between entities and their properties. However, this requires a laborious process and leaves the actual translation largely to manual coding. In contrast, our approach covers complex translations in terms of entire patterns and aims to automatically synthesise a suitable target representation and validate the translation.

Šváb Zamazal & Scharffe (2009) proposes a transformation service capable of generating alternative modelling choices while retaining the intended meaning of each transformed pattern. In contrast, we are concerned with translating information expressed in multiple formal languages and not to transform entire ontologies. More importantly, the authors assume that corresponding patterns are already given, or can easily be identified from purely structural criteria in the source and target ontologies.

Currently most state-of-the-art ontology matching techniques focus on the matching of "atomic" concepts and do not address translation between multi-entity concrete representations which require considering context (Ritze et al. 2009). These techniques are typically unable to resolve ontological heterogeneities between atomic concepts and a target ontology where the intended meaning is defined using multiple concepts or concepts with one or more attributes or relationships. Ritze et al. (2009) investigate matching of complex correspondences by means of patterns comprised of an atomic concept and attribute.

In contrast, our work focuses on larger fragments comprised of multiple elements, which may not be represented in the target ontology or may be represented differently. Therefore, our aim is to find and translate between corresponding complex templates. Further, our work characterises the key properties of templates into a "fragment ontology" and investigates the composition of fragments to synthesise complete translations into multiple different representations. Template translations can then be used as a semantic foundation for formal mapping-based translation tools as described in Berger et al. (2010).

Hamdi et al. (2010) are concerned with identifying refactorings in ontologies based on change patterns, where refactoring operations and implied mappings are derived from lexical similarity and structural subsumption. Although our aim is to transfer information rather than refactor ontologies, some of their pattern predicates may serve as inspiration for our ontology.

The patterns defined in Scharffe (2009) form a rather generic classification of correspondences (attributes, classes, and relations). The work seems mostly concerned with representing the heterogeneities as patterns. The focus is on relatively simple "patterns" in concept hierarchies without comprehensive domain axiomatisation.

5 Conclusion

In this paper we have described a potential solution to address problems arising from the need to transfer information between multiple ontology-supported yet incompatible information systems. Given comprehensive heterogeneities in data models and meaning alike, current approaches to schema- and ontology-matching are unable to overcome all difficulties.

We proposed an approach based on a unified ontology of "information fragments" that represent abstractions of concrete representations of chunks of information in different data models. Including associated translation operators between primitive data types, our model can support automated planning techniques in order to derive transformation operations from one representation into another. The same framework will also allow us to identify aspects of data models that cannot be translated adequately, hence aiding the developer in revising or extending the translation ontology and/or the underlying standards and their ontologies.

There are multiple avenues to follow-up on our preliminary investigations. First, ontology matching tools will need to be evaluated to assess which are most suitable to infer (mis-)alignment between existing reference data libraries of ISO 15926 and MIMOSA. As a result, an aligned unified ontology comprising the entities and relationships of both standards will be built. Second, the templates defined in the ISO 15926 data model will need to be complemented with templates sourced from the MIMOSA CCOM data model. Both will be formalised within the unified ontology. Third, planning and rewriting techniques will be investigated in order to create a reasoning framework that is effective in synthesising and validating translations as well as being scalable to real problems requiring incremental translation between multiple endpoints.

References

Berger, S., Grossmann, G., Stumptner, M. & Schrefl, M. (2010), Metamodel-Based Information Integration at Industrial Scale, *in* 'ACM/IEEE International Conference on Model Driven Engineering Languages and Systems (MoDELS)', Vol. LNCS 6395, Springer-Verlag, pp. 153–167.

Felfernig, A., Friedrich, G., Jannach, D., Stumptner, M. & Zanker, M. (2003), 'Configuration knowledge representation for semantic web applications', *Artificial Intelligence for Engineering Design, Analysis and Manufacturing* **17**(1), 31–50. Special issue on Configuration.

Hamdi, F., Reynaud, C. & Safar, B. (2010), Pattern-based mapping refinement, *in* P. Cimiano & H. Pinto, eds, 'Knowledge Engineering and Management by the Masses', Vol. 6317 of *Lecture Notes in Computer Science*, Springer Berlin / Heidelberg, pp. 1–15. 10.1007/978-3-642-16438-5_1.

ISO (2003), *Industrial Automation Systems and Integration – Integration of Lifecycle Data for Process Plants Including Oil and Gas Production Facilities – Part 2: Data Model*, International Standards Organisation, Geneva, Switzerland. Ref. No. ISO15926-2:2003(E).

ISO (2004), *Industrial Automation Systems and Integration – Integration of Lifecycle Data for Process Plants Including Oil and Gas Production Facilities – Part 1: Overview and fundamental principles*, International Standards Organisation, Geneva, Switzerland. Ref. No. ISO15926-1:2004(E).

Ritze, D., Meilicke, C., Šváb Zamazal, O. & Stuckenschmidt, H. (2009), A pattern-based ontology matching approach for detecting complex correspondences, *in* 'Proc. of Int. Workshop on Ontology Matching (OM)'.

Scharffe, F. (2009), Correspondence Patterns Representation, PhD thesis, University of Innsbruck.

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

Šváb Zamazal, O. & Scharffe, F. (2009), Pattern-based ontology transformation service, *in* 'Proceedings of the 1st International Conference on Knowledge Engineering and Ontology Development'.