Metadata Standards for Semantic Interoperability in Electronic Government

Jim Davies, Steve Harris, Charles Crichton, Aadya Shukla, and Jeremy Gibbons
Software Engineering Programme, University of Oxford
Wolfson Building, Parks Road, Oxford OX1 3QD, UK
Jim.Davies@comlab.ox.ac.uk

ABSTRACT

Effective data sharing, across government agencies and other organisations, relies upon agreed meanings and representations. A key, technological challenge in electronic governance is to ensure that the meaning of data items is accurately recorded, and accessible in an economical—effectively, automatic—fashion. In response, a variety of data and metadata standards have been put forward: from government departments, from industry groups, and from organisations such as the ISO and W3C.

This paper shows how the leading standard for metadata registration—ISO 11179—can be deployed without the need for a single, monolithic conceptualisation of the domain, and hence without the need for universal agreement upon a particular model of electronic governance. The advantages of this approach are discussed with regard to the UK eGovernment Interoperability Framework (eGIF) and the UK Integrated Public Sector Vocabulary (IPSV).

1. INTRODUCTION

Standards can have an important role to play in software development: by adhering to a published standard, we might hope that our applications or data would prove immediately compatible with those produced by others: for example, if we use tags only as allowed by an earlier HTML standard, we might be confident that our web pages would render successfully in a wide range of browsers.

In industry, there may be a competitive advantage in *not* following a standard: the benefits of compliance can be evaluated in purely commercial terms, and it may well be that they do not justify the costs. Suppliers will often add new features, adopt different interpretations, or omit certain aspects of a standard to reduce costs, to improve performance, or to make it more difficult for customers to use products from other suppliers.

In electronic government, the situation is reversed: it is often more important to work to agreed standards than to reduce costs, improve performance, or to seek to establish a greater degree of ownership or control over the development. Standards are the means by which electronic government can achieve interoperability across departments and agencies, improve their management of supplier contracts, and ensure that key data remains accessible over time.

Standardisation activity in software was originally focussed upon language and protocol design: upon the intended interpretation of programming statements, and upon the concrete representation of data and commands. Since then, there has been a pronounced shift in focus towards metadata standards: descriptions of intended functionality and meaning that can be associated with particular items of data, in order to ensure a consistent treatment and interpretation.

Initial work in this area was motivated by the concerns of document management: the widely-used Dublin Core standard, for example, is a collection of metadata items addressing issues such as the authorship, format, intended audience, and availability of information resources. Subsequent work has been focussed upon the data within documents, most notably, within the messages sent between different systems in business [5] and healthcare [8].

The importance of metadata standards at the level of individual data elements has already become apparent in these and other domains: for example,

- The NASA Mars Climate Orbiter [11] was lost after messages between two different systems were misinterpreted: the first was sending values in US Customary units (lbf-s), the second assumed that the values arriving were measured in SI units (Ns); the result was an initial orbit 170km lower than planned—23km below survivable height.
- US Government compliance data on water quality near industrial waste sites had to be re-evaluated when it was discovered that ‘rate of flow’ had been taken to mean the rate of flow of waste by some teams, while others had taken it to mean the rate of flow of the waterway [7].

Although in either case, misunderstandings *could* have been discovered earlier—and remedied—through improved communications, review and management activities, this kind of expensive mistake is hard to avoid when data is processed and integrated automatically, and its semantic consistency—the compatibility of units, and of intended interpretation—is checked only manually.

The need to record and promote—and where possible, automate—the re-use of standard metadata elements across enterprises and initiatives has led to the establishment of metadata *registries*, for example:

- the *Intelligent Transport Systems* metadata registry, a repository of data definitions and models for transport initiatives, currently being used by the UK Highways Agency to improve the quality of standards under development, such as the European DATEX II specification for travel information [15];
- the *National Information Exchange Model (NIEM)* maintained by the US Departments of Justice and Homeland Security is designed to "develop, disseminate and support enterprise-wide information exchange standards and processes... to effectively share critical information in emergency situations, as well as support the day-to-day operations of agencies throughout the nation" [6];
- the *METadata Online Registry (METeOR)* maintained by the Australian Institute of Health and Welfare, part of the Australian Government, is a repository of national data standards on health, housing and community services [12].

Each of the registries mentioned above is designed according to the model provided by ISO 11179 [9], an international standard for metadata registration. It is this standard, and its interpretation within the domain of electronic government, that is the subject of this paper.

The paper begins with an introduction to ISO 11179, together with an explanation of its relationship to existing electronic government interoperability frameworks. In Section 3, we identify a number of problems that may arise in the implementation of the standard and related initiatives—particularly in regard to the implied adoption of a single, conceptual domain.

In Section 4, we explain how these problems can be solved with a specific interpretation of the ISO 11179 standard, together with an extension of the metadata registry approach to include the registration of multiple models of definition, classification, and usage. The result is an approach to data and metadata standards that supports multiple perspectives on information and its meaning, an essential prerequisite for interoperability across different departments, agencies, and cultures. The paper ends with a brief discussion of the value of the interpretation and extension, particularly with respect to initiatives such as the UK electronic Government Interoperability Framework.

2. ISO 11179

The ISO/IEC 11179 standard *Metadata registries (MDR)* addresses "the semantics of data, the representation of data, and the registration of the descriptions of that data". Its purpose is to promote: standard descriptions, common understanding, harmonisation, and re-use of data in different contexts; and the management and re-use of the "components of data" [9]. It exists in six parts: the first of these describes the overall approach; the others address the specific concerns of classification, attributes and relationships, definitions, naming, and registration. Each of these has a bearing on the present discussion.

The approach to data semantics taken in the standard is characterised by the notion of *data element*, characterised by a combination of:

- a *data element concept*, corresponding to a *property* of an *object class*;
- a *value domain*, corresponding to a set of values that may be assigned to this concept, described in terms of a *conceptual domain*.

The notions of property and object class are only loosely defined: it is left to the implementer to decide upon their precise interpretation.

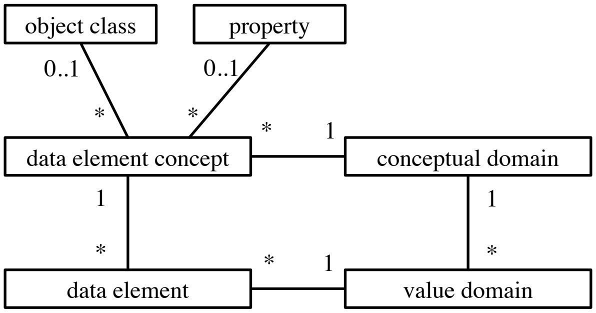

Figure 1: Basic entities in ISO 11179

The six different entities mentioned above are related according to the class diagram shown in Figure 1, based upon an entity-relationship diagram given in Part 3 of the standard. Note that a data element is uniquely characterised by the combination of a data element concept and a value domain.

Many data elements may share the same data element concept. This allows for different kinds of measurement of the same property: for example, the data element concept of a person's weight with two different value domains to produce two different data elements: a person's weight in kilograms and a person's weight in pounds. Similarly, many data element concepts may share the same conceptual domain, indicating that they share the same dimensionality: for example, that they are all measurements of mass.

It would seem logical for the association between data element concepts and conceptual domains to correspond to the composition of three other associations: that is, a data element concept is associated with a conceptual domain precisely when this is the conceptual domain for one of the value domains of an associated data element. However, the standard does not explicitly require that this is the case.

All of the above entities are maintained as administered metadata items within a registry. Each item is associated with a collection of administrative metadata, whose function is to support the processes of registration, maintenance, and re-use. For example, each item has an administration record, recording status, change information, and dates of creation, change, and effect. Guidelines for the registration and maintenance of administered items are set out in Part 6 of the standard.

In addition to these, a registry will store one or more classification schemes for administered items: each classification scheme is itself an administered item, associated with the same administrative metadata, and subject to the same procedures. Part 2 of the standard outlines the role of classification schemes, but leaves their specific nature to be determined by others: specifically, those working on related standards for concept systems, taxonomies, and ontologies. It is recognised, however, that a data element concept may appear in more than one classification scheme.

The two entities—object class and property—used to characterise a data element concept also have another role: the names of these entities can be used in the construction of names for administered items. Part 5 of the standard explains how this should be done, giving the example of a data element called

```
Cost Budget Period Total Amount
```

where `Cost` is an object class, `Total Amount` is a property, and `Budget Period` is an additional, *qualifying term*. `Amount` might also appear as a *representation term*, although its use would be redundant here. Names are associated with *contexts*—groupings of administered items.

The use of an underlying terminology for naming adds semantics to the administered items, over and above the explicit semantics of actual data definitions, outlined by Part 4 of the standard. Further, implicit semantic information may be derived from associations between different data element concepts, and between different conceptual domains: the standard allows both kinds of relationship to be recorded directly in the registry.

3. PROBLEMS

There is no *fundamental* problem with the ISO 11179 standard—it has no features or requirements that would make it unsuitable for use in electronic government. However, there are problems of interpretation and scope: there are features that may be mis-used, and features that are mising; as a result, an implementation may fail to achieve the desired level of data interoperability across different departments, agencies, and cultures. The features that may be mis-used are:

- the classification of data element concepts according to object class and property;
- the option to add direct associations between data element concepts, and between conceptual domains;
- the ability to update the definition of an administered item, and hence the semantics of existing data elements.

With regard to the specific requirements of electronic government, the missing features are:

- a precise, general notion of usage model, expanding upon the present notion of *context*, and encompassing optional patterns of data element usage, definition, and transformation;
- a standard means of extension: the standard acknowledges that each class will be extended in implementation, but provides no means of communicating the nature of that extension;
- support for the automatic integration or normalisation of data based upon data element definitions: there is no standard way of communicating that—in a particular context—one data element may be used in place of another, and how this may be achieved.

Before we propose any solution, we should explore exactly *why* these two sets of features—or rather, their possible misuse, and their absence, respectively—may present problems in the context of electronic government.

3.1 Single perspective

Although the standard allows for multiple classification schemes, the implication is that these do not add to the semantics of the administered items. Instead, the entirety of the semantic information consists in:

- the association of data element concepts with object classes and properties (and their relationship in any underlying terminology);
- any specified associations between data element concepts, and between conceptual domains;
- the textual descriptions associated with data element concepts, with non-enumerated value domains, and with each value of an enumerated value domain.

Two alternative perspectives upon semantics are afforded by Parts 3 and 4 of the standard. In the first part, greater emphasis is placed upon the use of object classes and properties; in the second, it is suggested that the ‘essential meaning’ of a concept should be determined entirely by its textual definition, without reference to other elements.

The second perspective is unrealistic for data elements of any subtlety or complexity: the meaning of a concept may be most obviously defined relative to that of others; if the same information is included independently in several definitions, effective maintenance may be impossible; and, perhaps most importantly, the meaning of a concept may be partly determined by the current context.

But this does not mean that we should accept the first. If the associations—with other elements, object classes and properties—convey meaning, then each data element has a meaning determined at least partly from a single perspective, represented by the collection of associations in the registry, and any underlying terminology used for object classes and properties.

Such a single perspective is problematic:

- it is expensive and difficult to maintain, as any addition needs to be made in the context of the semantics given to all of the existing elements and the relationships between them;
- it does not support the simultaneous registration of multiple perspectives upon the same data element, or the incremental development of data element semantics;
- it requires a single, centralised process of registration, and implicitly a single registration authority, for any collection of data elements whose semantics may be related.

There is also the inevitable problem that as the scope and coverage of the registry increases, then either the number of related data elements will increase—requiring a greater degree of qualification or complexity in definition—or the compromises implicit in the re-use of the same data element in different contexts, with no additional semantics, will increase instead. Either way, the quality and utility of the semantics will degrade.

In the existing implementations of ISO 11179 where semantics are given through association, the data elements are indeed defined from a single perspective. With the exception of the NIEM, all of the registries are clearly the responsibility of a single agency, and the need for interoperability across agencies and departments—and thus the need for multiple perspectives upon semantics—has yet to become an issue.

The NIEM registry has modules for each branch of government, presenting packages of metadata elements to cover message requirements within or between agencies. Data elements regarded as essential for interoperability, such as those corresponding to address information, are shared between packages. Parts 4 and 5 of the standard are adopted, and strict constraints are placed upon the composition of schemas to avoid implicit associations.

Agencies are not required to adopt NIEM internally: the standard acknowledges that the same person or object might be seen from quite different perspectives in two different information systems. All that is required is agreement upon a minimal set of attributes, and that semantic consistency is maintained in the structure of data that crosses agency lines. Despite this, NIEM still promotes the perspective of a single agency across the whole of the registry.

3.2 Object classes and names

The problems of a single perspective may be amplified through the implementation of the example naming convention presented in Part 5 of the standard, in which a concept is given a name, and the greater part of its semantics, through the naming of an object class, a property, and a qualification. This was precisely the case for the example data element presented in Section 2:

```
Cost Budget Period Total Amount
```

where `Cost` and `Total Amount` are names of an object class and property, and the qualification `Budget Period` appears as the value of the `object_class_qualifier` attribute within the data element concept.

In practice, Part 5 presents principles by which naming conventions can be developed and describes an *example* naming convention; it does not require that names are constructed from controlled terms in this way, but merely includes this as an example of a convention. Nevertheless, the effect has been to encourage the adoption of names that systematically encode the semantics of the element as expressed by the associations and the underlying terminology.

That this convention should have been adopted is perhaps surprising in the light of the following passage, taken from Part 1 of the standard:

> *It is important to distinguish an actual object class or property from its name. This is the distinction between concepts and their designations. Object classes and properties are concepts; their names are designations. Complications arise because people convey concepts through words (designations), and it is easy to confuse a concept with the designation used to represent it.*

The reduction of data element semantics to a structured collection of terms, even if those terms come from a controlled vocabulary, is unlikely to produce satisfactory results.

As an example of the difficulty of achieving a single perspective, consider the different ways in which the concept of 'proposed date of travel' might be given meaning through properties of object classes: to an immigration service, it might be a property of a visa application; to a security service or transport agency, it might be a property of an airline ticket, or of an individual.

The same confusion will apply whenever an attribute may correspond to an association between classes, or whenever there is no obvious candidate for the 'object' or 'class' in question: consider, for example, the measurement of the size of a hole—is this a property of the hole (seen as an object), or of the object in which the hole appears? These questions are easily answered within a specific context, where the conceptualisation has a particular purpose, but are impossible to answer in the abstract [13].

Even when the two or more agencies might agree upon the same choice of object class and property, agreement upon further qualification might be problematic, particularly when there may be a conflict of interests involved in the choice of definition. An ambulance service may wish to have *time of patient admission* be the time at which the patient is delivered to a medical facility; a hospital might wish to have the same data element record the time at which the patient is first seen by a doctor.

3.3 Context

In practice, the meaning of a data element is determined largely by the specific context of its application, even down to the level of a part of a specific form. For example, consider the 'date of birth' attribute upon the *Electronic Travel Authority* [1] that must be obtained in advance of air travel to Australia. To the immigration service, once the application is processed, this is a single data item. On the form however, the traveller is asked to enter their date of birth on two successive pages: that these are two different data items becomes apparent if a different date is entered, for the applicant is then flagged as a 'person of relevance', and the electronic application is rejected.

Every form, every database, every message, every usage of data provides a different context, and a potentially different semantics for the data element in question. Furthermore, this semantics may be determined collectively for a combination of data elements: the meaning of a measurement of geographical location may be explained, along with a street name and a postal code, by the context in which these items appear together. Without any way of representing this context as a model of usage, we are forced to resort to assigning meanings through the use of lengthy, textual descriptions, or a complex, problematic terminology.

The extent of the problem is illustrated by the difficulty of representing the data element `Person Birth Date` from the UK eGIF [4], which consists of two items: one for the actual date, and another for the evidence supporting the claim. This could be represented in ISO 11179 with a 'value domain relationship', using appropriate local semantics to compose a compound value domain and associating this with the appropriate data element, or one could use a 'data element derivation' with similar semantics to derive the compound data element from separately declared data elements. It would thus be possible for two implementations to represent even this data element in incompatible ways.

Another consideration is the way in which the usage, and thus the semantics, of data elements changes over time. The ISO 11179 standard includes provision for change management in data element definitions, but an obvious question remains: how do we know whether a new version of the definition is consistent with all existing usages of a data element? And even if this is the case, how is the information to be communicated to the present and future users of the data element in question?

The expectation of multiple, changing perspectives makes support for the automatic transformation of data, based upon semantic information, an essential requirement. We should be able to assert, within a specific context, that one data element is a suitable replacement for another, or record how data recorded against one collection of elements can be transformed to fit another. We should recognise also that these assertions and transformations provide additional, specific semantics to the data elements involved.

Finally, just as we might expect to see changes in the usage of data elements, we might also expect to see changes in the usage of a metadata registry. Although the standard allows extension of metadata entities, it does not provide for an extension in the functionality of a registry: in particular, there is no way for a registry to record and communicate the nature of any additional entities that it holds; as a result, any new model or pattern of usage of a data element must be introduced separately to each implementation.

4. SOLUTION

The problems described above can be solved through extension and specific interpretation of the ISO 11179 standard. We extend the notion of administered items to include an abstract representation of usage models, covering not only forms, schemas, and development models—in languages such as XForms, XML, and UML—but also classifications, ontologies, and transformations—in languages such as SKOS, OWL, and XSLT.

We extend the notion of a data element—the basic, representational entity, not a data element concept—to include references to multiple defining concepts and usages, with the intention that every usage of the element (and hence every implicit extension to its semantics) should be recorded.

At the same time, we constrain the interpretation of the standard in the following ways:

- data elements and value domains are given meanings only in the context of models: that is, concepts are defined only within models; through the use of object classes and properties, or otherwise;
- once a model—and hence a piece of semantic information—has been made available to users of the registry, it is no longer changeable;
- successive versions of a model exist as separate administered items; assertions of interoperability between them, and definitions of data transformations, are stored as models;
- no associations are maintained between administered items in the registry; instead, models may refer to each other, and to components, through unique identifiers and namespace conventions.

This results in a significant simplification of the core diagram, from the six classes of Figure 1 to the three classes of Figure 2: the conceptual classes have disappeared inside individual models, and there may be more than one of them—more than one semantic extension—per concrete data element.

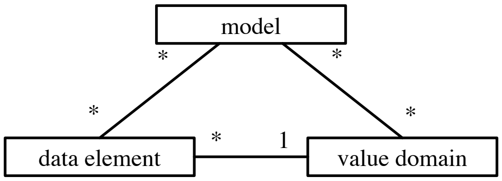

Figure 2: Data elements and value domains

We can address the problem of extensibility with a sufficiently general definition of models: for example, by defining a registry model as anything that can be given a metamodel using a standard modelling framework, such as the Eclipse Modeling Framework (EMF) [3], an open standard for model and metamodel representation.

Provided that each registry implementation had at least one model built in—the EMF metamodel—these metamodels could themselves be registered as models, and passed from registry to registry to announce and distribute support for new models of usage. Other likely candidates for built-in metamodel instances are shown in the extension of the model class in Figure 3.

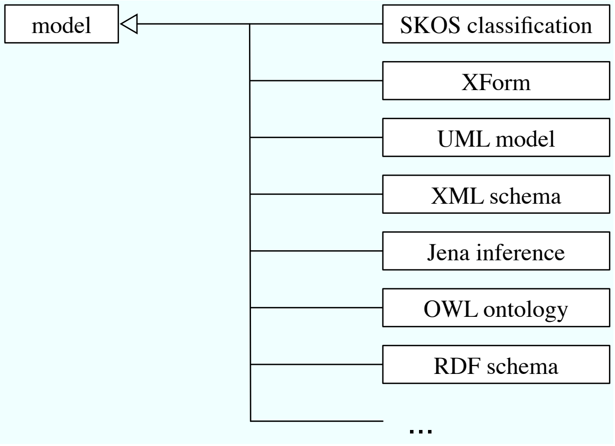

Figure 3: Extending models

With this approach, classification schemes are simply models to be registered, and can be adopted or combined as required: for example, we might wish to register the following simple fragment of classification shown in Figure 4, taken from the UK Integrated Public Service Vocabulary (IPSV) [4]. This is a SKOS classification asserting that, in the context of the IPSV, 'Rivers', 'Water Conservation', and 'Water Quality' are subcategories of 'Water Resources', which is itself a subcategory of 'Environment', related to 'Natural Habitats'.

A registry may contain several classification schemes, and several mappings between them: for example, in the UK we

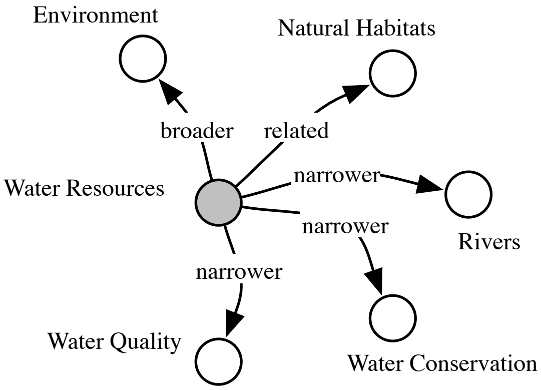

Figure 4: A simple classification

might expect a metadata registry for electronic government to include both the IPSV, which is defined from an internal communications perspective, and the Local Government Navigation List (LGNL), an alternative taxonomy oriented more towards communication with the public.

Similarly, we might expect a registry, or a federation of registries, to maintain different representations for compound data elements, also represented as models. Consider, for example, the model of an *address* provided by the British Standard BS7666, included in the electronic Government Interoperability Framework (eGIF) [4]: a fragment of this, represented as an XML schema, is presented in Figure 5.

```
<xsd:complexType name="BSaddressStructure">
  <xsd:sequence>
    <xsd:element name="SAON" type="AONstructure" ...
    <xsd:element name="PAON" type="AONstructure" ...
    <xsd:sequence>
      <xsd:element name="StreetDescription" ...
      <xsd:element
        name="UniqueStreetReferenceNumber"
        type="USRNtype" minOccurs="0" ...
    </xsd:sequence>
    <xsd:choice>
    ...
      <xsd:sequence>
        <xsd:element ref="Town"/>
        <xsd:element ref="AdministrativeArea" ...
      </xsd:sequence>
      <xsd:element ref="AdministrativeArea"/>
    </xsd:choice>
    <xsd:element name="PostTown" ...
    <xsd:element name="PostCode" ...
```

Figure 5: BS7666 address schema

The eGIF standard for addressing includes information of particular value to UK organisations: for example, the formula for establishing the validity of a UK postal code. Other countries, or other organisations, will find it more convenient to use other standards for addresses. We might expect a metadata registry—or a federation of registries—to hold several different standards, together with transformations between them.

The XSLT document in Figure 6 can be used by an XSLT processor to transform a BS7666 address into the OASIS standard used by the Australian government.

If this transformation is made available as an administered item, then users of a registry can apply it to transform address data collected against eGIF into the OASIS format, as shown in Figures 7 and 8.

```
...
<xsl:template match="b:BS7666Address">
  <xsl:element name="a:Address">
    ...
    <a:AdministrativeArea a:Type="State">
      <a:NameElement>
        <xsl:value-of select="b:PostTown"/>
      </a:NameElement>
    </a:AdministrativeArea>
    ...
    <a:Thoroughfare>
      <a:Number a:Type="Name">
        <xsl:value-of select="b:SAON/b:Description"/>
      </a:Number>
      <a:NameElement a:NameType="NameAndType">
        <xsl:value-of select="b:StreetDescription">
        </xsl:value-of>
      </a:NameElement>
    </a:Thoroughfare>
    ...
```

Figure 6: Address transformation

```
<BS7666Address>
  <SAON>
    <Description>Highways Agency ...
  </SAON>
  <PAON>
    <Description>123</Description>
  </PAON>
  <StreetDescription>
     Buckingham Palace Road
  </StreetDescription>
  <Locality> </Locality>
  <PostTown>
    London
  </PostTown>
  <PostCode>
    SW1W 9HA
  </PostCode>
</BS7666Address>
```

Figure 7: An address in BS7666

We would expect such transformations to be maintained by different agencies, just as the one of the local government bodies in the UK—Lichfield Council—is presently maintaining a transformation from the IPSV to the LGNL. These transformations would be versioned and combined in the same way as any other model stored in the registry.

The challenge of versioning, and of automatic transformation between data sets collected against related models, can be addressed through the use of specific ontologies. With appropriate references to other models, data elements, and transformations, these ontologies would specify whether and how data might be automatically integrated—even if this required that a 'lowest common denominator' be found for some of the compound data elements or structures.

For example, we might register and maintain an ontology asserting that data collected against model `C` can be mapped into the form specified in model `B`, that data collected against `B` or `D` can be mapped into the form specified in `A`. A textual version of the OWL model for this ontology is shown in Figure 9: extended with references to the actual models and transformations, this could be used to support the automatic

```
<Address>
  <AdministrativeArea Type="State">
    <NameElement>London</NameElement>
  </AdministrativeArea>
  <Thoroughfare>
    <Number>
      123
    </Number>
    <NameElement NameType="NameAndType">
      Buckingham Palace Road
    </NameElement>
  </Thoroughfare>
  <Premises Type="Building">
    <NameElement NameType="Name">
      Highways Agency
    </NameElement>
  </Premises>
  <PostCode>
    <Identifier Type="Number">SW1W 9HA</Identifier>
  </PostCode>
</Address>
```

Figure 8: An address in OASIS xAL

integration of data sets collected against these four models, using `A` as a common representation.

```
  Declaration(OWLClass(data-model))
  Declaration(ObjectProperty(transforms-to))
  TransitiveObjectProperty(transforms-to)
  ...
  ObjectPropertyDomain(transforms-to data-model)
  ObjectPropertyRange(transforms-to data-model)

  ...
  ObjectPropertyAssertion(transforms-to
    data-model-B data-model-A)
  ObjectPropertyAssertion(transforms-to
    data-model-C data-model-B)
  ObjectPropertyAssertion(transforms-to
    data-model-D data-model-A)
```

Figure 9: A transformation ontology

The same approach can be adopted to the updating of semantic definitions. Where the definition of a metadata item is updated—that is, when a new version is introduced into the registry—we can create or update an ontology that tells us whether the two versions are interoperable. If the item is a data model, then the ontology may explain how, and in what context, data collected against one version can be used with the other. An ontology can also be used to explain which version is preferred, and it may be that different versions are preferred in different contexts: the development of (successive versions of) a metadata item may fork, with different versions being maintained by different agencies.

5. DISCUSSION

5.1 Application to eGIF

Like many other interoperability frameworks, the UK eGIF is defined using the W3C XML schema language. Our interpretation of the ISO 11179 standard facilitates the automatic incorporation of such frameworks: instead of creating value domains for the XML data types declared, such as

```
  RestrictedStringType
```

we can include and refer to them in a single model, automatically generated from the framework schemas.

The lack of existing support in ISO 11179 for compound data elements would make it difficult to register eGIF items in a systematic, maintainable fashion: we would need to represent data elements of `xs:complexType`—of which there are many in eGIF—using a derivation association, resulting in the introduction of a single perspective (see Section 3.1) that would conflict with, for example, the LGNL taxonomy. The declaration of compound elements as structures within models avoids this problem.

To test the applicability of the approach, we translated the eGIF schema definitions—data elements and value domains—into administered data items within an ISO 11179 implementation, preserving the relationships with the IPSV vocabulary. Some technical issues were encountered:

– the use of inheritance, and the union of `simpleType`s, required an explicit model of value domain relationships;

– several of the `simpleType` facets used had no equivalent in the standard, and new representations were required;

– the use of anonymous simple types, a common feature in compound data element definitions, meant that names needed to be generated for each of the corresponding model elements.

More importantly, there was a significant difference in the use and interpretation of namespaces:

– in ISO 11179, namespaces are used to indicate registration authorities or stewardship, depending upon whether a registry is maintained by a single agency, or shared amongst several;

– in eGIF, namespaces are assigned according to purpose or context, and are used within schemas to qualify or disambiguate common terms in the vocabulary.

The ISO 11179 approach is to be preferred, for otherwise namespaces can act in the same way as associations between metadata items in the registry, adding unadministered semantics from a single perspective; although they are an important tool at both the modelling and the programming level, they need to be aligned with registered (collections of) models if we are to achieve the expected levels of transparency and interoperability.

The XML schema language has recently been extended with support for semantic annotation, using Semantic Annotations for the Web Services Description Language WSDL (SAWSDL) [10]: in combination with OWL and SKOS, this allows the representation of semantics without any need to consider an additional standard.

The transformation exercise, however, provided further evidence as to the value of the ISO 11179 standard approach. Without the additional level of organisation, it is possible to add nodes and comments in the schema declarations without adhering to any documented structure or systematic approach; there are places within the eGIF framework where semantic information is inconsistently represented, making interoperability and automatic processing difficult to achieve and maintain.

5.2 Universality

Despite the issues raised in Section 3, an international effort exists to establish a *universal* data element framework: the UDEF [14]. This includes 16 basic object classes and 18 properties against which data elements may be modelled. Each basic object and property has a textual description, which is further qualified to define subclasses.

In UDEF, subclasses have no informal (textual) definition of their own, but derive their semantics from their relationship to the superclass. The intended scope should be clear from the inclusion of the following:

```
UDEF:ap.15 Anti-Matter.Substance
UDEF:a.c.p.9 Operating.System.Software.Product
UDEF:c.6 Mathematical.Law-Rule
```

The utility of such a framework is questionable:

- it raises questions such as whether 'antimatter' should be categorised as a form of 'substance', where the answer clearly depends upon purpose, context, or even scientific progress;
- it is difficult to imagine a lawyer, a software engineer, or a physical scientist accepting the UDEF taxonomy as a useful categorisation, when a city government law, a mathematical law, and a physical law belong to the same category.

Furthermore, UDEF may be seen as an *upper ontology* of data elements. Attempts to derive such ontologies have been largely abandoned because "different conceptualisations which serve as inputs to ontology are likely to be not only of widely differing quality but also mutually inconsistent" [2].

A more practical approach would be for the UDEF to define a partial metadata registry and register objects and properties according to the standard. As it stands, it is difficult to recommend this framework to architects and developers in the electronic government domain.

5.3 Distributed, open semantics

It is useful to consider the differences between the approach set out in this paper and the work of the eXtended MetaData Registry (XMDR) project, which shares many of the same objectives, being "concerned with the development of improved standards and technology for storing and retrieving the semantics of data elements, terminologies, and concept structures in metadata registries" [17].

This output of this work includes an implementation of the ISO 11179 metamodel in OWL, producing a single concept system in which the semantics of data elements are given in terms of object classes and properties, and thus enforcing a single perspective. In contrast, our approach allows the introduction of multiple, possibly inconsistent, classifications and ontologies involving the same data elements: any semantics, inference, or transformation is relative to a particular context, described by a combination of models.

The importance of simultaneous support for multiple perspectives, whether they reflect those of different communities of interest, or that of a single community, evolving over time, cannot be overstated: without this support, we should expect that an interoperability framework or initiative will

- become increasingly expensive to maintain as the size of the registry increases;
- act as an inhibitor for process change and innovation in government;
- fail to deliver the expected levels of data sharing and semantic interoperability.

To insist upon the exclusive use of such a resource could compound the problem: "taking away people's ability to manage their own information in their own way could significantly reduce government performance at all levels" [16].

Conversely, providing simultaneous support for alternative perspectives, captured in collections of models, opens up opportunities for the distributed development and maintenance of semantic resources. What is more, it provides a platform for *open semantics*: to extent that they are published within a standard registry, the meanings, intentions, and usages of data elements, value domains, ontologies, and models are accessible to all.

6. REFERENCES

[1] Australian Department of Immigration and Citizenship. Electronic Travel Authority System. `http://www.eta.immi.gov.au/`.

[2] Barry Smith. Ontology. In *The Blackwell Guide to the Philosophy of Computing and Information*. Blackwell, 2004.

[3] Frank Budinsky, David Steinberg, Ed Merks, Raymond Ellersick, and Timothy J. Grose. *Eclipse Modeling Framework*. Prentice Hall, 2003.

[4] Stella Dextre Clarke. e-GIF, e-GMS and IPSV: What's in it for us? *Legal Information Management*, 7(04):275–277, 2007.

[5] cXML. commerce eXtensible Markup Language. `http://www.cxml.org/`.

[6] Department of Justice and Department of Homeland Security. National Information Exchange Model. `http://www.niem.gov/`.

[7] Larry Fitzwater. Plenary Address: Standards Overview. 11th Metadata Open Forum, 2008.

[8] HL7. Health Level Seven. `http://www.hl7.org/`.

[9] International Organization for Standardization. ISO 11179: Information Technology Specification and Standardization of Data Elements. `http://www.iso.org/`.

[10] Jacek Kopecký, Tomas Vitvar, Carine Bournez, and Joel Farrell. Semantic Annotations for WSDL and XML Schema. *IEEE Internet Computing*, 2007.

[11] NASA. Mars Climate Orbiter Mishap Investigation Board Phase I Report, 1999.

[12] Australian Institute of Health and Welfare. Metadata online registry. `http://meteor.aihw.gov.au/`.

[13] Barry Smith, Werner Ceusters, and Rita Temmerman. Wüsteria. In *Medical Informatics Europe*, 2005.

[14] The Open Group. Universal Data Element Framework (UDEF). `www.opengroup.org/udef`.

[15] UK Highways Agency. Transport Systems Metadata Registry. `http://www.itsregistry.org.uk/`.

[16] Jan Wyllie. The Integrated Public Service Vocabulary: A Confusion of Poly-Hierarchies. *Enterprise Information*, 2, 2005.

[17] The eXtended MetaData Registry (XMDR) Project. `http://www.xmdr.org/`.