Faculty of
Computer Science

universität
wien

Technical Report TR-20070901 December 1999

No Author

Putting the CIDOC CRM into Practice
Experiences and Challenges

Philipp Nussbaumer[1] and Bernhard Haslhofer[2]

[1] Research Studios, Studio Digital Memory Engineering, Vienna, Austria
philipp.nussbaumer@researchstudio.at

[2] University of Vienna, Department of Distributed and Multimedia Systems
bernhard.haslhofer@univie.ac.at

Abstract. In the course of the BRICKS project we had to provide uniform access to archaeological metadata stored in various autonomous and distributed repositories. A potential solution to the immanent problem of establishing metadata interoperability, lies in the utilisation of the CIDOC CRM, a global ontology which has gained much attention in the cultural heritage domain. However, the CIDOC CRM constitutes only a formal, semantic specification and abstracts from any implementation issues. This gap between the well-defined conceptual and the undefined technical level could result in additional incompatibility instead of interoperability. In this paper, we point out the experiences and challenges we have encountered while integrating metadata from archaeological institutions using the CIDOC CRM. With this work we aim to share best practices, point out shortcomings, and provide input for a scientific discourse.

1 Introduction

In the cultural heritage field institutions such as museums, archives, or libraries are facing a growing need to integrate their system with those of other institutions. The goal of most integration projects is to provide uniform access to the digital assets stored in a set of distributed, autonomous, and institution-specific repositories. This is also the case for the scenario we are focusing on in this paper: in the course of the BRICKS project [1], one goal was to provide uniform access to the digital assets of several archaeological institutions[3]. The digital assets we were dealing with are findings, or more specifically, coins that were excavated in the UK and filed in distinct, institution-specific systems, using proprietary cataloguing standards.

Thus, we were facing the very well known problem of information integration — gaining interoperability among heterogeneous, proprietary systems, or more specifically, among the metadata stored therein. There are several ways of how interoperability on the metadata level can be achieved [2], the introduction of a

[3] e.g. the Portable Antiquities Scheme (http://www.finds.org.uk), and the Archaeology Data Service (http://ads.ahds.ac.uk)

global conceptual model[4] being one of them. Such a model formalises the notions and defines the concepts for a certain domain (e.g. cultural heritage, archaeology). If all institutional repositories express their existing metadata according to this model, it is possible to gain uniform access by formulating queries over a single ontology. For the cultural heritage domain, the CIDOC CRM (CIDOC Conceptual Reference Model [3]) provides such a global model or — as it is called in its specification — ontology.

As for any integration scenario which is based on the usage of a global ontology, integrating metadata from various sources using the CIDOC CRM involves three main steps: the first step — *mapping* — embraces the alignment of the data sources' semantic and structural definitions (i.e. their schemes) with those of the CIDOC CRM. In a second step — *lifting and normalisation* — the data and schemes must be lifted to a common technical representation. This is a prerequisite to enable queries over the data sources using a certain query language. *Data processing* in a certain application context forms the third and final step. Intuitive interfaces are needed that hide the complexity of the underlying global ontology from the end user.

In this paper, we describe our methodology of utilising the CIDOC CRM for integrating metadata from various autonomous and distributed archaeological institutions. Since the BRICKS infrastructure [4] is largely based on Semantic Web technologies (RDF [5], OWL [6]), these build the foundation for a technical realisation of the methodology. After describing the requirements in Section 2, we propose a methodology for the previously identified steps (Section 3). We explain the details of this methodology in Section 4, shortly present a prototypical implementation in Section 5, give an overview of the related work in Section 6, and finally discuss the experiences and problems in Section 7.

2 Requirements

There are various ways of achieving interoperability in metadata integration scenarios (as further elaborated in Section 6). Integrating data while leaving the existing source systems unchanged, can be accomplished by agreement upon a global ontology onto which the source data is mapped. Other techniques, by contrast, may demand rather ample changes in the source systems, e.g. when imposing a standardised application-specific schema that all systems have to adopt.

In our context, one of the tasks was to integrate metadata and content of various archaeological institutions in order to allow uniform access to their assets; hence, making them query-able via a single search interface. Since the involved institutions already had their systems in place and did not have an incentive to adapt them according to any specific requirements, imposing a standardized schema (as described above) or other similar "intrusive" techniques was not an

[4] Different domains use different terms such as 'model', 'schema', 'ontology', and 'vocabulary' to denote models representing real-world entities. In this paper we use these notions interchangeably.

option. Within our scenario an end-user application had to be implemented, enabling both domain-experts as well as non-professionals to identify archaeological findings by searching and browsing the involved institutions' reference collections of items. As it is in line with both the institutions' as well as the technical requirements (some of which are described below), a global ontology approach for data integration was agreed on.

2.1 Mapping

Applying a global ontology approach for achieving interoperability requires each source schema to be mapped against the concepts defined in that ontology.

The issues that arise when local schemes are mapped against global ontologies are:

- *Discovering mapping relations*
  Identifying semantic correspondences between the concepts of the global and those of the local schemes requires experts that precisely know the semantic definitions of both. Although there exist many semi-automatic approaches that could assist the experts in discovering the mappings (e.g. [7, 8]), intellectual effort will always be required to determine the correctness of a mapping.
- *Representing mappings*
  After having discovered the mappings, the experts must specify or represent them in a machine processable way so that the system can take them as input for further data processing steps. The formalism for representing mappings must be strong enough to precisely capture the details of any semantic correspondence.
- *Decentralisation of metadata schema mapping*
  Parties mapping their data to the global ontology should be able to do so in a preferably decentralised manner, i.e. without the need of aligning their mappings with other parties, thus with minimal need of central coordination.

In our setting, the source schemes to be integrated were initially limited to collections of coins of the Portable Antiquities Scheme (PAS) and the Archaeology Data Service (ADS), both of which correspond to simple metadata schemes modelled as elementary attribute-value pairs. Figure 1 shows the fields of both schemes and depicts their semantical and syntactical correspondences.

The schemes are very similar in structure and semantics — most of the fields are identical (semantically and even syntactically in terms of their XML representation), whereas a few represent a semantically equivalent concept only.

Integration of these two data sets seems rather trivial — indeed, as most of the fields are equivalent, the ADS schema could be directly mapped to the PAS schema without any loss of semantics. In doing so, the PAS schema would serve as the global ontology. However, as straightforward an integration of these two metadata schemes may seem, complexity is likely to increase when additional data sources and schemes are introduced. Since the application-specific PAS

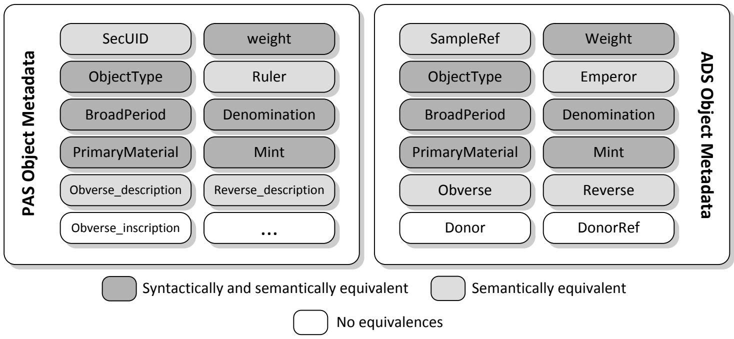


Fig. 1. Correspondences between the PAS and ADS metadata fields

metadata schema is proprietary and not very expressive, more complex schemes may not be mapped against it without considerable loss of semantics. Hence, a global ontology used for data integration has to be sufficiently expressive to cover the semantics of all involved source data schemes.

2.2 Lifting and Normalisation

Accessing digital assets in distributed repositories requires the uniform representation of the metadata and schema definitions in all involved data sources. This is because uniform access is typically provided via a certain query language, which in turn is bound to a certain data model. SQL, for instance, is bound to the relational model, while SPARQL works for RDF graphs.

Usually, this task is performed by a wrapper component, which is built on top of an existing system. A wrapper rewrites queries posed in the integration system's query language to queries that can be executed at a data source and returns the resulting data in a uniform data format. However, in the BRICKS setting the institutions have not opened the query interfaces to their native systems, but have exposed their metadata via the Open Archives Initiative Protocol for Metadata Harvesting[5] or provided their metadata as XML dumps of their databases.

2.3 Data Processing

The use of a global ontology can substantially reduce the complexity of search and retrieval at the client application. The application may access the data uniformly and does not need to include knowledge of all source schemes but only of the global ontology.

However, a global ontology itself can be very complex. Since the end users will formulate queries over this ontology and it is unlikely that they will do

[5] http://www.openarchives.org/OAI/openarchivesprotocol.html

that by writing queries in a certain query language (e.g. SQL, SPARQL), it is necessary to hide this complexity from the user and provide an easy-to-use and yet powerful Graphical User Interface (GUI) for efficient search and retrieval.

2.4 General Requirements

Since the previously mentioned requirements (e.g. mapping) were not only conceptual ones but also had to be implemented into a running system, it is clear that their realisation had a strong dependency on the technical environment imposed by the BRICKS framework. There, all metadata schemes — also any global ontology — are expressed in OWL DL and metadata are represented in RDF. This implies that mappings had to be defined to a global ontology expressed in OWL DL, metadata exposed via OAI-PMH (or XML dumps) had to be lifted to RDF, and SPARQL served as query language for formulating queries over the global ontology.

Another issue having a strong impact on all these requirements is the nature of the global ontology. In our application context, we chose to use the CIDOC Conceptual Reference Model (CRM) which is aiming at the provision of a global schema, embracing the semantics of concepts used in the field of cultural heritage. Because the CRM is by definition not bound to any technology, we chose its OWL definition[6] for our implementation.

3 Semantic Integration of Archaeological Findings using the CIDOC CRM

Application of a global ontology in data integration scenarios has obvious benefits of leaving existing systems unchanged and easing data storage, search and retrieval. Even so, there are some general issues that have to be undertaken when applying a global ontology approach — below we will discuss these issues and propose a methodology of how to address them, taking the CIDOC CRM as an example ontology. We believe, however, that this methodology is not restricted to the use with the CIDOC CRM but holds for other global ontologies and application scenarios as well.

3.1 A CIDOC CRM Integration Methodology

The CIDOC CRM provides no guidance in mapping schemes to the model or encoding, storing, or processing mapped data. Yet these issues are central when implementing the model in an application. We have identified three main issues for adopting the CRM:

- **Mapping**: As an initial step, the source schemes have to be mapped to the global ontology. The mappings have to be provided by domain experts, respectively experts of the source and target schema. Since multiple schemes

[6] `http://cidoc.ics.forth.gr/OWL/cidoc_v4.2.owl`

may be mapped to the global ontology in parallel by different institutions, mechanisms are needed to handle different mappings for semantically equivalent metadata as well as issues arising from identical mappings for semantically differing metadata.

- **Data Lifting and Normalisation**: Having mapped the metadata schemes to the global ontology, the institutions' instance data must be made available to the application, thereby lifted and normalised into a common representation. Depending on the technical infrastructure this can be achieved by wrapping existing data sources or by transforming existing data and ingesting them into a target system that is capable of answering queries in a specific query language.

- **Data Processing**: Global ontologies embrace a multitude of modelling primitives of metadata schemes used in a specific domain, and thus are rather complex in terms of syntax and semantics. The complexity of the underlying schema is consequently reflected in the complexity of queries for search and retrieval. In end-user applications it might therefore be necessary to hide this complexity from the user, leaving the client-application with the issue of user interfaces transparently operating on the global ontology.

3.2 The CIDOC CRM

The CIDOC Conceptual Reference Model is an object-oriented ontology for the domain of cultural heritage. It has been developed to meet the needs of integrating, mediating and exchanging information from museums, libraries and archives. Version 4.2.1 of the ontology consists of a set of 81 classes and 132 properties to describe things, concepts, people, places and time and their relationships, thus enabling the creation of information networks. The main purpose of the CRM is to provide the semantic basis to describe data models and metadata schemes already in use within the cultural heritage domain.

The model offers solely high-level conceptions of how to describe entities and their relations. It does not present a methodology or guidance to what should be documented nor responds to application-specific consistency or implementation issues (regarding e.g. data typing).

All classes of the CRM — excluding class *Primitive Value* and its sub-classes — are (direct or indirect) sub-classes of class *CRM Entity*, which comprises all things that may be described using the reference model. Figure 2 shows a small section of the CRM class hierarchy relevant for the examples presented in this paper. Aforementioned sub-classes of class *Primitive Value* (namely *Number*, *Time Primitive* and *String*) are not considered as elements within the universe of discourse and are not further elaborated within the model. Similar to the modelling classes of the ontology, these primitives abstract from any specific implementation in terms of storage and processing.

Analogous to the data model of the Resource Description Framework (RDF), the building blocks of the CIDOC CRM are *triples* (entity–property–entity),

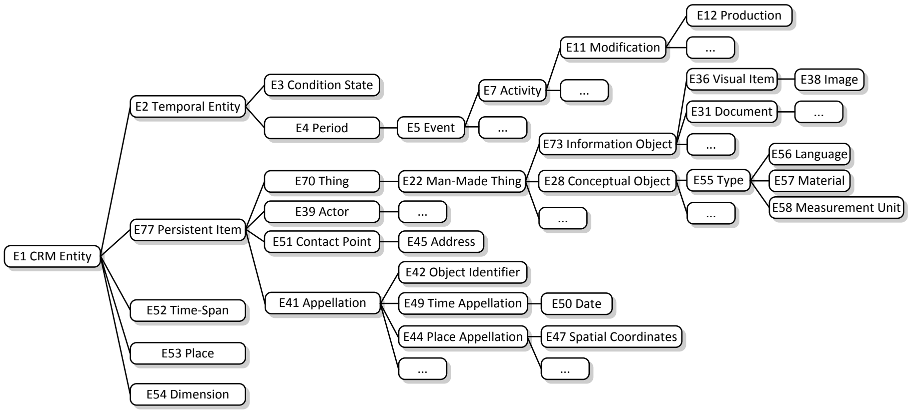
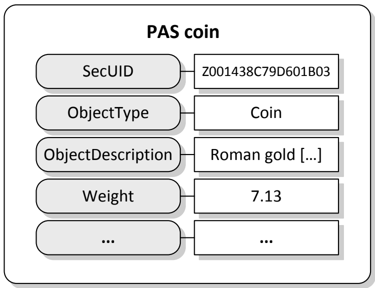


Fig. 2. Small section of the CIDOC CRM class hierarchy (version 4.2.1)

where properties are used to describe relationships between entities. The defined properties' domains and ranges are restricted to specific classes. This means that properties are restricted to relate entities of fixed classes, for example, property *P2 has type* has a domain of *E1 CRM Entity* and a range of *E55 Type*. Properties may not only be used to relate entities but also to relate classes to properties, i.e. properties may themselves have properties.

The CRM offers both strict and multiple inheritance for classes and properties; since the classes of the model are sub-classes of *E1*, property *P2*, for example, may be applied to each (non-primitive) class in the model. Using its inheritance mechanisms, the model may be extended by sub-classing existing concepts. By this means the model may be easily complemented by application-specific taxonomies or controlled vocabularies in general (e.g. by extending class *E55 Type*).



PAS coin

SecUID	Z001438C79D601B03
ObjectType	Coin
ObjectDescription	Roman gold [...]
Weight	7.13
...	...

Fig. 3. A coin described using the PAS metadata schema

In the previous section we have discussed the PAS metadata schema which consists of attribute–value pairs. A part of a coin instance described using the PAS scheme is depicted in Figure 3, whereas Figure 4 shows an example of how the same instance may be represented using the CRM.

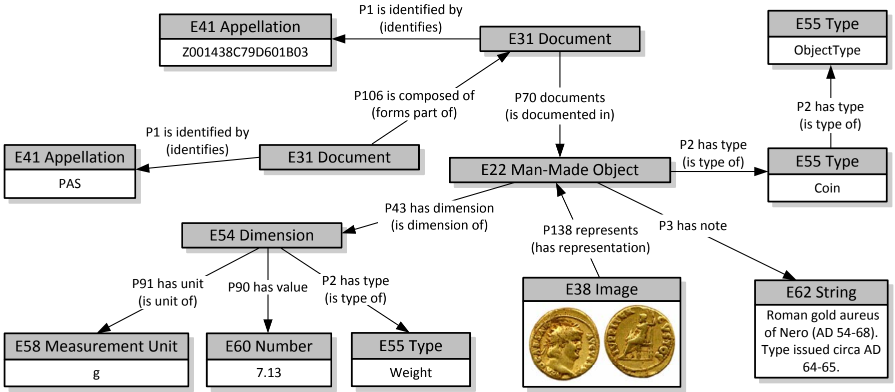


Fig. 4. A coin from the Portable Antiquities Scheme described using the CIDOC CRM

Both the CRM and PAS schema representations contain the same set of metadata. The directed CRM graph, however, provides slightly extended semantics of the modelled metadata, the main facts depicted being:

- The described object is a *Man-Made Object* of type "Coin", whereas its *Type* is declared as "ObjectType".
- It *is documented in* a *Document identified by* an *Appellation* "Z0014[...]", forming part of another *Document identified by* the *Appellation* "PAS".
- The coin has a documented *Dimension* of *Type* "Weight", having the *Number* value "7.13" and a *Measurement Unit* of "g", as well as a *String* note of "Roman gold aureus [...]".

Note that the CRM graph also expresses metadata that is only implicitly available in the original data (see Figure 3), such as the weight dimension or a measurement unit[7].

The CIDOC CRM concepts and related properties provide only high-level semantics in terms of concepts that might be used in the cultural heritage domain. In the example we have introduced the field names of the PAS schema as application-specific vocabulary for *E55 type*s. This vocabulary is needed for providing means of identifying (and therefore querying and retrieving) the "metadata attributes" represented as sub-graphs. Lacking such a vocabulary, the se-

[7] These information are not encoded in the source system's item instance data and have been added for illustrative purpose.

mantics of the CIDOC CRM's classes and properties does not suffice to e.g. describe the different dimensions of a coin: the statements of “Dimension has value” and “Dimension has unit” are ambiguous when describing multiple dimensions sharing the same unit (e.g. diameter and thickness). Adding the sub-statement “Dimension has type”, containing an according term of the used vocabulary, allows stating various dimensions unambiguously.

It has been mentioned before that the CIDOC CRM deliberately omits implementation guidelines for its model. For practical applications, the model has been encoded in RDF/S[8] and OWL. In general, the OWL representation of the CIDOC CRM represents properties as *ObjectProperties*, therefore allowing their use for relating classes only. To relate entities and literals, OWL prescribes the use of *DatatypeProperties* — such properties have to be introduced to the model to enable and restrict use of primitive data values (like string, integer, float etc.) within the graph.

The next section elaborates on how we applied the proposed methodology when implementing an end-user application aiming at the integration of different data sources from archaeological institutions, as discussed above.

4 Details

Applying the CIDOC CRM for our application's purpose required three steps and their associated issues to be mastered: mapping the source data to the ontology, importing it into the system, and processing (search, retrieve and render) the data at run-time. The following sections will discuss each individual step in more detail and offer exemplary solutions to common problems that might arise.

4.1 Step 1: Mapping the source schemes to the CIDOC CRM

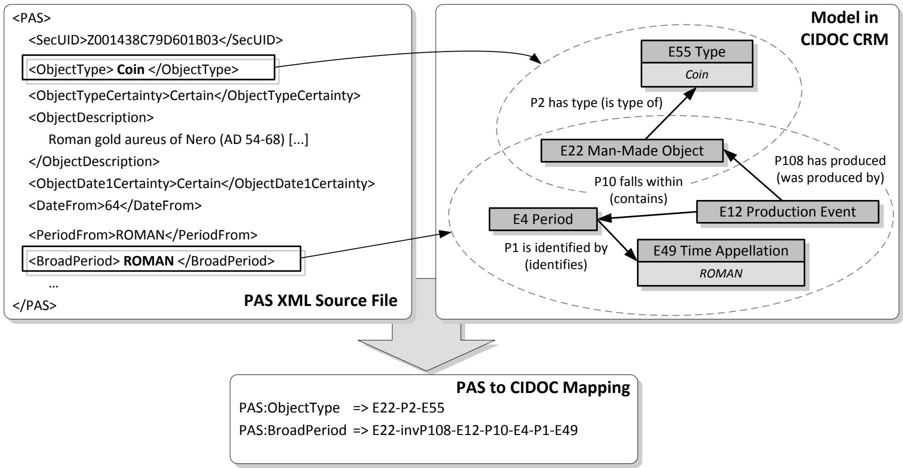
The central idea of obtaining metadata interoperability using the CIDOC CRM is to map each proprietary schema to a global ontology.

In a first step, an expert familiar with the source schema has to identify which metadata attributes to map. The mapping definitions are then represented by unambiguous *mapping chains* (or *paths*), indicating a sequence of entities and properties of the CIDOC CRM. Figure 5 gives an example attribute mapping of the PAS metadata schema.

The attribute-value pairs of the source instance are mapped to corresponding sub-graphs in the CRM representation. As indicated in Figure 5, these sub-graphs may be alternatively represented by the chains of the corresponding entities and properties. Hence the mapping specification may be reduced to the assignment of such chains to the attributes of the source schema.

The CIDOC CRM provides no guidance for the domain or schema experts which metadata attributes of the source schema to map, and which classes and

[8] `http://cidoc.ics.forth.gr/rdfs/cidoc_v4.2.rdfs`



```
<PAS>
  <SecUID>Z001438C79D601B03</SecUID>
  <ObjectType> Coin </ObjectType>
  <ObjectTypeCertainty>Certain</ObjectTypeCertainty>
  <ObjectDescription>
    Roman gold aureus of Nero (AD 54-68) [...]
  </ObjectDescription>
  <ObjectDate1Certainty>Certain</ObjectDate1Certainty>
  <DateFrom>64</DateFrom>
  <PeriodFrom>ROMAN</PeriodFrom>
  <BroadPeriod> ROMAN </BroadPeriod>
    ...
</PAS>
```

```
PAS:ObjectType   => E22-P2-E55
PAS:BroadPeriod  => E22-invP108-E12-P10-E4-P1-E49
```

Fig. 5. Mapping PAS schema elements to the CIDOC CRM

properties should be used in doing so. Apart from the intellectual challenge of assigning meaningful and valid chains to individual metadata information of the source schema, the mapping process already has to take into account some implementation issues. Since the model does not imply how or where to store literal values in the graph, extensions of the model and their proper use must be clarified initially.

Defining only entities and relations between entities the CIDOC CRM leaves the question where to store actual data values open to the application developer. We have defined properties (owl:DatatypeProperties) for being able to store actual values (e.g. a string "coin"). Each property has the respective XML Schema data types [9] as range and all classes of the model as domain. So every class in the graph may have an associated literal value[9].

Mapping source schemes to the global ontology is likely to be a decentralised process, each institution's source schema being mapped by a respective expert. When mapping schemes to the CRM, it is therefore most likely that semantically equivalent metadata attributes are mapped to different chains or identical chains are applied for semantically different metadata attributes. Consider the following examples for each of the possible mapping inconsistencies:

- *Different chains for equivalent metadata*
  As depicted in Fig. 1, the metadata fields "SecUID" (PAS) and "SampleRef" (ADS) are semantically equivalent, denoting an (internal) item identifier. The mappings below describe semantically equivalent metadata using differ-

[9] This might seem counterintuitive, since the CRM already defines classes of primitive values like *E62 String* or *E60 Number*. However, these classes are restricted to be used with few properties, so restricting the DatatypeProperties to the primitive value classes is not sufficient.

ent chains:

```
PAS:SecUID       E22-P47-E42
ADS:SampleRef    E22-invP70-E31-P1-E41
```

While the PAS identifier (PAS:SecUID) is encoded as "(E22) *Man-Made Object* (P47) *is identified by* (E42) *Object Identifier*", the semantically equivalent ADS identifier (ADS:SampleRef) is represented by "(E22) *Man-Made Object* (invP70)[10] *is documented in* (E31) *Document* (P1) *is identified by* (E41) *Appellation*". As a result, the semantic equivalence fails to be preserved in the target schema. In our application we have introduced an internal vocabulary which defines concepts for grouping mappings with the same semantics. For instance the chains depicted above may internally be mapped to the concept "identifier". Due to the introduction of such a vocabulary, feedback cycles between the mapping experts and the application administrator(s) are required.

– *Identical chains for different metadata*
The CIDOC CRM classes and properties provide means of describing information with rather high-level semantics. Therefore similar concepts may be expressed using identical chains. The following example shows the mapping of easting and northing coordinates of the PAS schema:

```
PAS:northing   E22-P53-E53-P87-E47
PAS:easting    E22-P53-E53-P87-E47
```

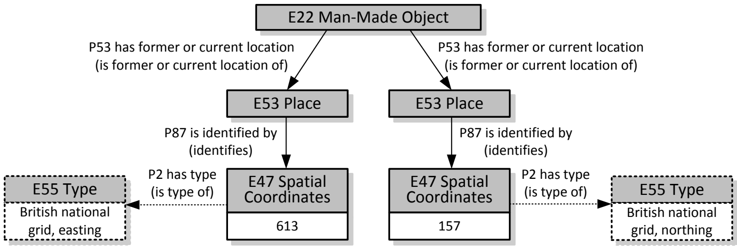

E22 Man-Made Object
P53 has former or current location (is former or current location of)
P53 has former or current location (is former or current location of)
E53 Place
E53 Place
P87 is identified by (identifies)
P87 is identified by (identifies)
E55 Type
British national grid, easting
P2 has type (is type of)
E47 Spatial Coordinates
613
E47 Spatial Coordinates
157
P2 has type (is type of)
E55 Type
British national grid, northing

Fig. 6. Adding assertion chains to uniquely identify sub-graphs

If such ambiguities appear in the very same schema mapping, they may be resolved by adding additional *assertion* chains, uniquely identifying a subgraph:

[10] *inv*: abbr. for *inverse*, meaning that the property's semantics changes from "documents" to "is documented in".

```
PAS:northing   E22-P53-E53-P87-E47
                               E47-P2-E55 (northing)
PAS:easting    E22-P53-E53-P87-E47
                               E47-P2-E55 (easting)
```

The values denoting northing and easting coordinates may now be stored as string value appended to entity *E47 Spatial Coordinates* in a *E53 Place* sub-graph each, uniquely identified by the assertion chains containing fixed values (depicted in parentheses, see also Figure 6 for the resulting graph). Note, though in this example the assertion values contain meaningful information themselves, they are needed to uniquely identify the sub-graphs representing the northing and easting values.

Ambiguities spanning over different schema mappings may only be resolved by a central instance, e.g. the application administrator.

Depending on how the mapping specification is encoded, it might be used as direct input or has to be processed itself in an intermediate step to be applicable in the process of transforming the source data into their target representation. In our application context, we have provided spreadsheets to the mapping experts, enabling them to specify the mappings in a form analogous to the "*metadata attribute – CRM chain*" notation above. The spreadsheets are semi-automatically transformed into XSL stylesheets that are then used to transform the source data into the RDF/XML target representation at ingestion time (see section 4.2).

Since the CRM's high-level semantics may require enrichment of the mapped metadata by controlled vocabularies to preserve the original semantics and/or to uniquely identify the metadata information, it is evident that querying and retrieving the data involves not only expert knowledge of the CIDOC CRM but also of the source schema mappings as well as the employed vocabularies. As the semantics of the metadata and their encoding chains (or sub-graphs) are available in the mapping specification, we decided to re-use this information in the application itself to configure an easy-to-use graphical query interface (see section 4.3).

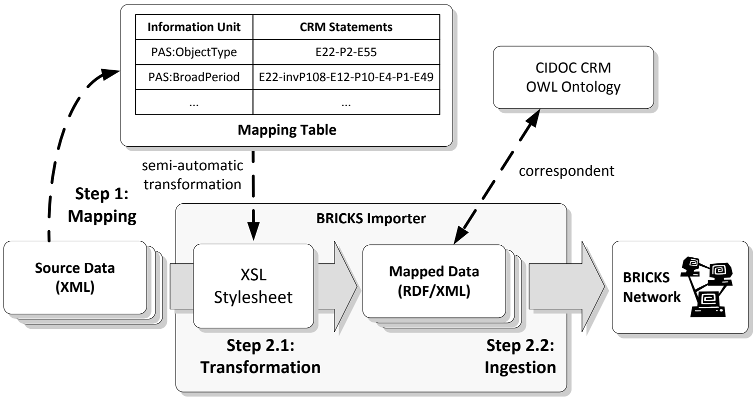
4.2 Step 2: Lifting and Normalization

In general, there are two ways of lifting and normalising data in an integration context. Wrapper components encapsulate the source systems and mediate between the source scheme and the global scheme at request-time. Another possibility lies in transforming and ingesting the data into a target system, thereby replicating it.

Responding to the wish of the participating institutions to leave their existing systems unchanged, the BRICKS framework [10] has been designed to *import* (and therefore replicate) data from source systems, rather than accessing the participating institutions' existing systems using wrapper components.

Using a replication approach, the process of lifting and normalising the source data involves two individual steps, namely (i) the *data transformation* according to the mapping specification created in Step 1 and (ii) the actual *data ingestion* whereby the (transformed) data is stored in the system.

In our application's context, we semi-automatically transform the mappings that are specified in a spreadsheet into an XSL stylesheet which is then used to transform the metadata. The BRICKS import infrastructure accomplishes transformation and ingestion of source data in a single combined step. By means of a dedicated importing application (BRICKS Importer [11]) the source data is transformed into RDF/XML (corresponding to the CIDOC CRM OWL representation) via the generated stylesheet and then stored into the BRICKS network infrastructure. Within the network the metadata is stored using a Jena[11] storage back-end. The basic workflow of this process is depicted in Figure 7.



Information Unit	CRM Statements
PAS:ObjectType	E22-P2-E55
PAS:BroadPeriod	E22-invP108-E12-P10-E4-P1-E49
...	...

Fig. 7. Importing (i.e. mapping, transforming, ingesting) metadata into BRICKS

4.3 Step 3: Processing the data at run-time

The steps discussed so far are concerned with aligning metadata with the global CIDOC CRM ontology and normalising (lifting) them to a uniform representation. The third and final step in our proposed methodology is concerned with providing means of searching, retrieving and rendering the integrated metadata.

Search and Retrieval Generally, the BRICKS network infrastructure exposes the stored data both via a full-text and a SPARQL query interface: a full-text search targets only the literal values encoded in the global ontology's instance data. This might not be satisfactory as it disregards the underlying schema's semantics. The SPARQL interface, in contrast, provides means to formulate

[11] `http://jena.sourceforge.net`

structured, semantically rich queries regarding the schema's semantics. The one query interface's strength proves to be the other's weakness: means of full-text search are easy to use, since the user needs to have little knowledge of the schema's semantics; a user interface expecting SPARQL queries, however, may overstrain users by demanding rather detailed knowledge of the target schema and the query language.

In case of the CIDOC CRM, formulating queries respecting the (semantical) correct combination of classes and properties of the model exhibits additional complexity. The CRM makes no propositions on which metadata information to map or which combination of classes and properties to use, hence there are a great many possibilities regarding the structure of the actual graph that represents the mapped metadata. Consequently, in application scenarios that involve multiple institutions mapping their metadata independently from each other, semantically equivalent metadata might be encoded using different structures and vocabularies. Querying for specific aspects (e.g. a coin's diameter) thus requires incorporation of mapping information, i.e. the different chains and vocabularies used when mapping the metadata to the CIDOC CRM.

Considering the need of an easy-to-use yet powerful graphical user interface, we decided to provide a configurable faceted-style search, taking into account both the semantics of the integrated data and the possibility of different mapping structures.

Our solution involves creation of SPARQL queries from the mapping chains known by means of the mapping specification created in Step 1. These chains are used to query along the "metadata attributes" encoded implicitly in the metadata graph, whereas the "attributes" refer to the source schemes' concepts initially mapped to the CIDOC CRM. For example, when querying for all coins from the Roman period, the mapping chains of the "period" attributes of both the ADS and PAS schemes have to be combined in a single query over the CIDOC CRM-encoded metadata.

In our application, we associate these metadata attributes to a set of controlled terms, i.e. we use an application-specific vocabulary to unambiguously identify their semantics. For example, as all metadata schemes might support metadata information equivalent to the notion of a coin's *diameter*, the source attributes' mappings may be grouped and associated to the term "diameter". Searching for specific aspects of an item is therefore possible by combination of the different mappings into a single query.

We use this approach in our application to provide a faceted search incorporating the aforementioned controlled terms (e.g. diameter, denomination, ruler) that are associated to the mapped metadata: the search interface has been designed to allow the user to answer several questions either by entering or selecting an according value as answer, whereas each provided answer narrows the result set and by this means supports the user in retrieving the designated information. The interface is presented along with a description of a prototypical implementation in Section 5.

Rendering When querying for specific aspects within the integrated data sources, we build combined queries from the mapping chains of the different source mappings. Rendering items of the result set poses the issue of how to retrieve all relevant metadata information and how to present it to the user. The CIDOC CRM's structure may suggest a presentation of the item metadata as a graph, so a user can browse through the metadata and the associated concepts. However, such a graph representation of the data — with nodes and edges labelled using the CIDOC CRM's conceptualisation terms — lacks comprehensibility. Understanding of the information in such a graph is further impeded by possibly inconsistent sub-graphs for equivalent metadata information, caused by irregular mappings of the source data.

Already utilising an application-specific vocabulary to associate equivalent metadata mappings to unambiguous terms, we decided to also use the mappings to extract the metadata information from the graph. To retrieve an item's metadata we therefore iterate the metadata mappings to extract the according information and render the result as attribute–value pairs (e.g. "diameter: 8.17"). Given that the metadata mappings are unambiguously associated to vocabulary terms, such a representation is semantically well-defined and easier to comprehend.

5 Prototypical Implementation

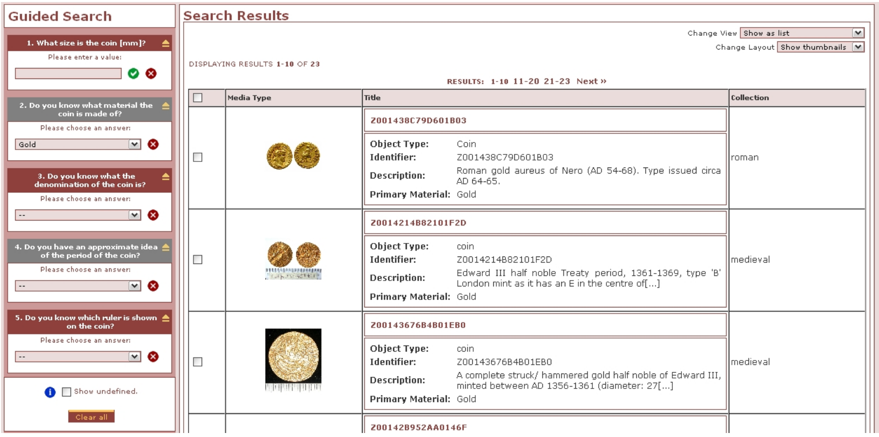
The Finds Identifier application, implemented using the Apache Struts framework[12], is a tool for expert users and non-professionals to identify findings made all over Europe. In our prototype implementation, the integrated findings are restricted to coins found in the United Kingdom.

A user may explore the reference collections in different ways in order to identify a finding:

- *Browsing* This functionality allows a user to browse the collections to get an understanding of the reference items' properties.
- *Simple Search* The Simple Search provides a full-text search interface, targeting the literal values of the available items' metadata instances.
- *Guided Search* Using the Guided Search, a user may answer several questions about prominent features of the finding. Thereby the user may easily find a representative set of similar findings.

Browsing and querying via full-text search are rather unsuitable for specific queries. When searching for Roman coins, for example, the full-text search returns not only the coins having defined a metadatum "period" having the value "Roman" but also coins where the term is contained in other metadata (like e.g. "description").
The Guided Search functionality (depicted in Figure 8) was introduced to allow such semantics-aware queries. Each question is transparently associated to an

[12] `http://struts.apache.org`



Guided Search

1. What size is the coin [mm]?
Please enter a value:

2. Do you know what material the coin is made of?
Please choose an answer: Gold

3. Do you know what the denomination of the coin is?
Please choose an answer: --

4. Do you have an approximate idea of the period of the coin?
Please choose an answer: --

5. Do you know which ruler is shown on the coin?
Please choose an answer: --

□ Show undefined.

Clear all

Search Results

Change View: Show as list
Change Layout: Show thumbnails

DISPLAYING RESULTS **1-10** OF **23**

RESULTS: 1-10 11-20 21-23 Next »

□	Media Type	Title	Collection
□		**Z001438C79D601B03** **Object Type:** Coin **Identifier:** Z001438C79D601B03 **Description:** Roman gold aureus of Nero (AD 54-68). Type issued circa AD 64-65. **Primary Material:** Gold	roman
□		**Z0014214B82101F2D** **Object Type:** coin **Identifier:** Z0014214B82101F2D **Description:** Edward III half noble Treaty period, 1361-1369, type 'B' London mint as it has an E in the centre of[...] **Primary Material:** Gold	medieval
□		**Z00143676B4B01EB0** **Object Type:** coin **Identifier:** Z00143676B4B01EB0 **Description:** A complete struck/ hammered gold half noble of Edward III, minted between AD 1356-1361 (diameter: 27[...] **Primary Material:** Gold	medieval
		Z00142B952AA0146F	

Fig. 8. Guided search interface

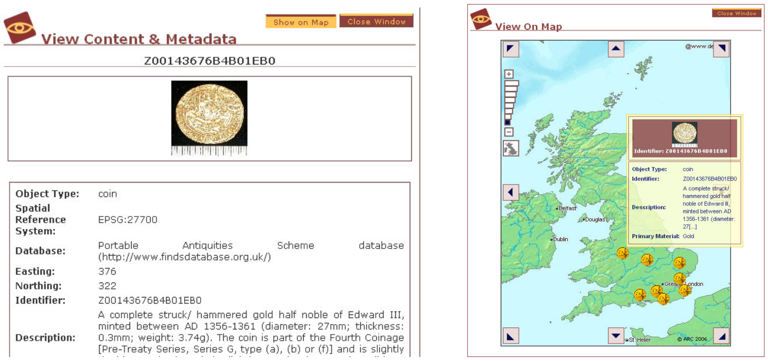
unambiguous term of the internal vocabulary, each of which groups all different mappings for a specific metadata information. The second question ("Do you know what material the coin is made of?"), for example, is associated to the term "material". When selecting an appropriate answer, the application creates a combined query which includes all mapping information grouped by the term "material", thus all items of the specific material — independent from the specific mappings — are returned.

To preclude queries that return empty result sets, the query interface is dynamically updated with every answered question. The lists of possible answers of the remaining questions are adapted to contain only values applicable to the current result set, helping to further narrow the set of similar findings.

For each item the user may then display the associated metadata (Figure 9(a)) as well as the finding place on a map (Figure 9(b)).

6 Related Work

The goal of the CIDOC CRM is to establish semantic interoperability among heterogeneous information sources in the cultural heritage domain. Considering the fact that an information source in this domain typically stores digital objects together with their metadata, and that metadata are the key for managing and accessing these digital objects [12], the goal is actually to establish semantic *metadata interoperability* among cultural heritage data sources. The CRM defines a global ontology against which the underlying semantics of local data base schemes or document structures can be mapped. However, such a *global ontology* approach is not the only way to achieve metadata interoperability. Other complementary techniques are listed in [2]: agreement on a single standard, creation of application profiles [13], and bilateral mappings between source schemes are some of them.



Z00143676B4B01EB0

Object Type:	coin
Spatial Reference System:	EPSG:27700
Database:	Portable Antiquities Scheme database (http://www.findsdatabase.org.uk/)
Easting:	376
Northing:	322
Identifier:	Z00143676B4B01EB0
Description:	A complete struck/ hammered gold half noble of Edward III, minted between AD 1356-1361 (diameter: 27mm; thickness: 0.3mm; weight: 3.74g). The coin is part of the Fourth Coinage [Pre-Treaty Series, Series G, type (a), (b) or (f)] and is slightly

Identifier: Z00143676B4B01EB0

Object Type:	coin
Identifier:	Z00143676B4B01EB0
Description:	A complete struck/ hammered gold half noble of Edward III, minted between AD 1356-1361 (diameter: 27[...]
Primary Material:	Gold

(a) Detail view of a coin's metadata

(b) Map view indicating the finding place of a coin

Fig. 9. Different views of a coin

If data sources are integrated using the global CIDOC CRM ontology, the data sources can be accessed by formulating queries over the global ontology's concepts. Basically there are two architectural possibilities for integrating the data sources: centralised or decentralised. In a centralised approach the metadata are converted according to the structural and semantic definitions in the CIDOC CRM and transferred into a central data store. In a decentralised approach the metadata reside in the data sources and are virtually integrated using a mediated query system or mediator-wrapper architecture [14]: the mediator exposes a mediation ontology — in this case the global CRM ontology — and provides a uniform query interface for the underlying data sources. Incoming CRM queries are unfold and forwarded to wrappers which encapsulate the local data sources and have the ability to answer those queries.

Having a single global ontology which provides a shared vocabulary for the data sources is different from multiple ontology approaches where each information source is described by its own ontology and semantic interoperability is established by bilateral mappings [15]. Examples for a distributed single ontology approach are the SIMS Information Mediator [16] and Ontobroker [17]. Observer [18], Edutella [19], and GridVine [20] are examples for multiple ontology approaches that employ inter-ontology mappings.

With its 81 classes and 132 properties in version 4.2.1, the CIDOC CRM is meant to be a very general, global ontology which formalises the notions in the cultural heritage domain. These kind of ontologies or models also exist in other domains: the Functional Requirements for Bibliographic Records (FRBR) [21] is an entity-relationsship model which should serve as a generalised view of the bibliographic universe, intended to be independent of any cataloguing standard or implementation [22]. The Suggested Upper Merged Ontology (SUMO)[13] is another example for a global ontology that "will promote data interoperability,

[13] http://ontology.teknowledge.com/

information search and retrieval, automated inferencing, and natural language processing" [23]. It defines high level concepts such as Object, ContinuousObject, Process, Quantity, a.s.o. The Descriptive Ontology for Linguistic and Cognitive Engineering (DOLCE) [24] is yet another example for a global ontology. Besides the global ontology, the authors also have specified a library of extensions which covers ontologies such as the Ontology of Description and Situations, an Ontology of Plans, and an Ontology of Web Services.

Noy [25] states that the communities do not yet have enough experience to claim that global ontology approaches are a success. She refers to two reports, one on the success [26] and another on the difficulties [27] of reusing global ontologies in integration scenarios. Wache [28] asserts that no global ontology can be defined in such a way that it fulfils all ontological requirements of all possible information systems that are integrated in a certain domain. Halevy et al [29] argue that in large scale environments a global ontology, which is the integrated part of a system that should facilitate information integration, becomes bottleneck in the process. It must be designed and maintained very carefully an cannot change significantly without violating existing mappings from data sources. In general, global ontologies only work well in integration scenarios where the sources to be integrated provide nearly the same view of a domain [15].

Mapping between ontologies and/or schemes can be represented in several ways. Having a mapping formalism that can capture the heterogeneities among the schemes is one approach. The MAFRA ontology mapping framework [30] is an example of a system that covers the whole heterogeneity spectrum through the definition of "semantic bridges". There exist several types of such bridges, each covering several heterogeneity problems — not forgetting the procedural information for "instance transformation". Another approach is to represent mappings as views or queries. Piazza [31] is a representative for such a system.

Metadata mappings are the technical specifications that serve as input for a process, which is commonly referred to as query reformulation. If we regard metadata mappings as a way of defining how to construct the elements of a target schema (e.g. the user selected schema) from the data in a source schema (e.g. a data source's schema), they fulfil the same functionality as views. Hence, if mapping specifications are not available in terms of views per se, they must be transformed into such a representation. In principle, there are two ways of representing mappings using views: (i) the data sources, i.e. their schema elements, are described as queries (views) over the user selected schema — this is referred to as *Local as View (LaV)* — or (ii) the user selected schema is described as a set of views over the data sources — this is known as *Global as View (GaV)*. The first case, query reformulation, means rewriting the queries similar to rewriting a query using a view [32]. In the second case, reformulation works analogously to view unfolding in traditional relational database systems.

7 Conclusions and Future Work

The goal of this paper is to narrow the gap between metadata interoperability on a conceptual level, as it is enforced by global ontologies such as the CIDOC CRM, and its technical realisation. We believe that the broader this gap is, the more likely it is that interoperability efforts could end up in technical incompatibilities. We have described a methodology consisting of three main steps (mapping, lifting and normalisation, data processing), each reflecting our experiences from implementing the CIDOC CRM in a real world setting.

In a data integration scenario, a global ontology provides the concepts against which the data source specific schema elements are mapped and over which user requests are formulated. For both tasks we have two main issues that could impede the actual goal of metadata interoperability.

The first issue is the abstractness of the concepts (e.g. Time Appellation, Man-Made Object) defined by the global ontology, which makes them ambiguous to any human user. Even expert users that are very familiar with the CIDOC CRM and the institution-specific schemes have produced ambiguous mappings and have required several iterations to produce consistent mapping definitions. If several experts specify mappings independently from each other, it is very likely that they will produce incompatible mappings and fail the goal of enabling interoperability.

Another point which is directly connected to the abstractness of the concepts, is the presentation to the user. For instance, to retrieve coins from a certain age (e.g. "Roman") from all available data sources, he or she must search for "E22 Man-Made Object – P108 was produced by – E12 Production Event – P10 falls within – E4 Period – P1 is identified by – E49 Time Appellation – *Roman*". Since this is not intuitive at all and not practical for the end users, it basically requires a graphical user interface which hides the complexity of the global ontology and allows the user to formulate queries over more concrete concepts. This could be selection boxes for the required item type (e.g. coin) or fields for specifying the item's attributes (e.g. period).

The second issue is the lack of technical specifications in global ontologies such as the CIDOC CRM. Without any detailed instructions of how to implement the mappings, represent instances, and process data during run-time, it is likely that each institution applies its own interpretation on a standardised global ontology. This again causes heterogeneities in scenarios that initially have aimed at providing interoperability. We therefore recommend to combine any attempt of providing conceptual interoperability with a detailed technical specification.

In this paper, we have discussed how we have dealt with these issues to allow uniform data access: mainly by using mapping information and internal vocabularies to be able to downright *restore* the original data semantics. We observe that this results in a major conceptual problem: the goal of using a global ontology in a data integration scenario is uniform, (source) application independent access — inclusion of the original mapping information and vocabularies somewhat re-establishes the original problem. Uniform access is hindered by the mere fact that information from the *source* data is needed to meaningfully query and interpret the *integrated* data.

To generalise and summarise our observations on using global ontologies, we refer to the fact that global ontologies are actually meant to serve as a reference for future ontologies that have not yet been developed. This however is not the case in integration scenarios where source schemes are already in place. As a result, they are not extensions of the global ontology but — from a semantical point of view — *artificially* mapped schemes. From that we can conclude that global ontology approaches, such as the CIDOC Conceptual Reference Model, might be suitable providing interoperability among data source that have not been implemented yet. However, in our integration context they proved to be unsuitable.

In the future, we will further elaborate on the mapping aspect of this work because we believe that the mapping capabilities of global ontology approaches (mapping by extension) are far too restricted for providing interoperability also on the technical level. We believe that experts must have the possibility to define technically precise mappings directly among their schemes and schemes of other data sources, without using a global ontology. Besides a formalism for representing and executing these mappings, this also requires mechanisms for determining their quality, for reacting on changes in the involved sources, and for keeping track of the available schemes and specified mappings.

8 Acknowledgments

The work reported in this paper was partly funded by the Austrian Federal Ministry of Economics and Labour and the European Union as part of the 6th Framework Program.

References

1. EU-FP6: BRICKS – Building Resources for Integrated Cultural Knowledge Services (IST 507457) (2007) Available at: `http://www.brickscommunity.org`.
2. Chan, L.M., Zeng, M.L.: Metadata Interoperability and Standardization – A Study of Methodology Part I + II. D-Lib Magazine **12**(6) (June 2006)
3. CIDOC Documentation Standards Group: CIDOC Conceptual Reference Model (CRM) – ISO 21127:2006. (December 2006)
4. Haslhofer, B., Hecht, R.: Metadata Management in a Heterogeneous Digital Library. In Cunningham, P., Cunningham, M., eds.: Innovation and the Knowledge Economy, eChallenges. Volume 2., IOS Press (2005) 967–973
5. W3C Semantic Web Activity – RDF Core Working Group: Resource Description Framework (RDF) (2004) Available at: `http://www.w3.org/RDF/`.
6. W3C Semantic Web Activity – Web Ontology Working Group: Web Ontology Language (OWL) (2004) Available at: `http://www.w3.org/2004/OWL/`.
7. Kalfoglou, Y., Schorlemmer, M.: Ontology Mapping: The State of the Art. Knowl. Eng. Rev. **18**(1) (2003) 1–31

8. Rahm, E., Bernstein, P.A.: A survey of approaches to automatic schema matching. The VLDB Journal **10**(4) (2001) 334–350
9. W3C: XML Schema 1.1 Part 1: Structure (August 2006) Available at: `http://www.w3.org/TR/xmlschema11-1/`.
10. Risse, T., Kneževic, P., Meghini, C., Basile, F.: The BRICKS Infrastructure — An Overview. In: The International Conference EVA2005, Moscow. (2005)
11. Hecht, R., Haslhofer, B.: Joining the BRICKS Network — A Piece of Cake. In: The International Conference EVA2005, Moscow. (2005)
12. Sheth, A., Klas, W.: Multimedia Data Management: Using Metadata to Integrate and Apply Digital Media Media. Mcgraw-Hill (1998)
13. Heery, R., Patel, M.: Application profiles: mixing and matching metadata schemas (September 2000) Available at: `http://www.ariadne.ac.uk/issue25/app-profiles/`.
14. Wiederhold, G.: Mediators in the Architecture of Future Information Systems. Computer **25**(3) (1992) 38–49
15. Wache, H., Vögele, T., Visser, U., Stuckenschmidt, H., Schuster, G., Neumann, H., Hübner, S.: Ontology-Based Integration of Information — A Survey of Existing Approaches. In Stuckenschmidt, H., ed.: IJCAI–01 Workshop: Ontologies and Information Sharing. (2001) 108–117
16. Arens, Y., Hsu, C.N., Knoblock, C.A.: Query Processing in the SIMS Information Mediator. In Huhns, M.N., Singh, M.P., eds.: Readings in Agents. Morgan Kaufmann, San Francisco, CA, USA (1997) 82–90
17. Ontobroker: Ontology Based Access to Distributed and Semi-Structured Information. In: DS-8: Proceedings of the IFIP TC2/WG2.6 Eighth Working Conference on Database Semantics-Semantic Issues in Multimedia Systems, Deventer, The Netherlands, The Netherlands, Kluwer, B.V. (1998)
18. Mena, E., Illarramendi, A., Kashyap, V., Sheth, A.P.: OBSERVER: An Approach for Query Processing in Global Information Systems Based on Interoperation Across Pre-Existing Ontologies. Distrib. Parallel Databases **8**(2) (2000) 223–271
19. Nejdl, W., Wolf, B., Qu, C., Decker, S., Sintek, M., Naeve, A., Nilsson, M., Palmér, M., Risch, T.: EDUTELLA: A P2P Networking Infrastructure Based on RDF. In: WWW '02: Proceedings of the 11th International Conference on World Wide Web, New York, NY, USA, ACM Press (2002) 604–615
20. Aberer, K., Cudre-Mauroux, P., Hauswirth, M., van Pelt, T.: GridVine: Building Internet-Scale Semantic Overlay Networks. In: International Semantic Web Conference (ISWC). Volume 3298 of LNCS. (2004) 107–121
21. IFLA Study Group on the Functional Requirements for Bibliographic Records, International Federation of Library Assocations: Functional Requirements for Bibliographic Records (September 1997) Available at: `http://www.ifla.org/VII/s13/frbr/frbr.htm`.
22. Tillett, B.: What is FRBR — A Conceptual Model for the Bibliographic Universe (2004) Available at: `http://www.loc.gov/cds/FRBR.html`.
23. Niles, I., Pease, A.: Towards a Standard Upper Ontology. In: FOIS '01: Proceedings of the international conference on Formal Ontology in Information Systems, New York, NY, USA, ACM Press (2001) 2–9
24. WonderWeb Consortium: DOLCE: A Descriptive Ontology for Linguistic and Cognitive Engineering (2003) Available at: `http://www.loa-cnr.it/DOLCE.html`.
25. Noy, N.F.: Semantic Integration: A Survey Of Ontology-Based Approaches. SIGMOD Rec. **33**(4) (2004) 65–70

26. Polyak, S.T., Lee, J., et al.: Applying the Process Interchange Format (PIF) to a Supply Chain Process Interoperability Scenario. In: Workshop on Applications of Ontologies and Problem Solving Methods, ECAI'98, Brighton, England (1998)
27. Valente, A., Russ, T., MacGregor, R., Swartout, W.: Building and (Re)Using an Ontology of Air Campaign Planning. IEEE Intelligent Systems **14**(1) (1999) 27–36
28. Wache, H.: Semantische Mediation für heterogene Informationsquellen. PhD thesis, University of Bremen (2003)
29. Halevy, Y., Ives, G., Suciu, D., Tatarinov, I.: Schema Mediation For Large-Scale Semantic Data Sharing. The VLDB Journal **14**(1) (2005) 68–83
30. Maedche, A., Motik, B., Silva, N., Volz, R.: MAFRA - An Ontology Mapping Framework in the Semantic Web. In: Proceedings of the ECAI Workshop on Knowledge Transformation, Lyon, France, 2002. (2002)
31. Halevy, A.Y., Ives, Z.G., Mork, P., Tatarinov, I.: Piazza: Data Management Infrastructure for Semantic Web Applications. In: WWW '03: Proceedings of the 12th international conference on World Wide Web, New York, NY, USA, ACM Press (2003) 556–567
32. Halevy, A.Y.: Answering Queries Using Views: A Survey. The VLDB Journal **10**(4) (2001) 270–294