

Archetype-based semantic mediation: Incremental provisioning of data services

Jesus Bisbal, Gerhard Engelbrecht, and Alejandro Frangi
CISTIB - Universitat Pompeu Fabra, and CIBER-BBN, Barcelona, Spain
{name}.{surname}@upf.edu

Abstract

Modern organizations need to exploit the information stored in heterogeneous and interrelated data sources, but often have no means to integrate them in a principled fashion. This general database research challenge is particularly relevant in distributed e-Science. Specifically, biomedical research generates a vast amount of heterogeneous data, which exceeds the current technological capacity to exploit it efficiently. Typically, service-oriented architectures are used in this context to define a unified view over all sources to be integrated. This unified schema needs to be mapped onto the underlying data sources, often including also semantic annotations. This approach suffers from high complexity and setup costs. In this paper we propose a novel application of semantic and mediation technologies, which leads to an incremental and on-demand definition of data mediation services. The so-called archetypes provide the context and semantics needed to setup such services, which significantly simplify their definition.

1 Introduction

The growing amount of heterogeneous and distributed information in e-Science infrastructures exceeds the current capacity of enterprise data management and data integration platforms, which necessitates new approaches to efficiently synthesize the available data [2]. Service-oriented architectures are usually employed to enable unified access to different information sources along with extensive descriptions of all data-sources, through often domain-specific ontologies. Many recent biomedical research projects [3, 7, 8, 11, 16] employed this approach, but also required huge efforts to adopt semantic technologies. On the one hand, ontological concepts need to be linked to actual data (*annotation*), and on the other hand, integration services must be defined based on semantic concepts (*semantic data mediation*). Both tasks are complex and time-consuming.

In this paper we propose an innovative application of so-called *archetypes* [1, 5] along with existing develop-

ments for data mediation [18] to enable incremental semantic data mediation in a service-oriented architecture. This archetype-based semantic mediation (ABSM) involves a novel linking of ontology concepts to actual data utilizing pre-existing archetypes and an automated generation of data services exposing related data sources in a mediated fashion. Thus, the overall effort for both the annotation work and the generation of data mediation services will be improved significantly. We even hypothesize that, based on the semantic annotations via archetypes, these data mediation services can be created automatically and enable incremental and on-demand (pay-as-you-go) semantic integration.

The following section reviews the related work. The ABSM approach is detailed in Section 3. Section 4 describes applications of this approach in the context of two EU-funded projects. The final sections conclude the paper and provide a number of future directions for this research.

2 State of the Art

This section describes relevant work on the core concepts and technologies used in this paper.

2.1 Archetypes

A novel approach to the modeling and management of information separates domain information models into three independent components: data, archetypes (i.e. structure), and semantics (i.e. ontology) [7]. Archetypes are used to define the different sets of related data attributes, which together provide the necessary context to adequately interpret the information which is to be stored, communicated or shared. External ontologies or terminological resources are optionally referred to from the archetype definitions [17] in order to annotate those definitions with semantically rich resources [19], facilitating interoperability and integration.

This approach originated in the medical informatics community [1, 5] to standardize medical information communication, but it is of general applicability. It advocates a principled design methodology to information modeling.

Archetypes are at the core of this methodology, which is often referred to as *two-level modeling* paradigm [5]. Its first level, the *reference model*, is a pre-defined set of very abstract classes and aggregation rules that provide the flexibility to model any information concept. Its second level, the *Archetypes*, add semantics and constraints to the potential instances of the reference model, in order to ensure that the desired semantics are captured by the clinical concepts defined by those archetypes.

The two-level modeling paradigm is being standardized by the major bodies in medical informatics (Europe's CEN 13606, and USA's HL7 RIM v3). While traditional information modeling methodologies produce large and detailed domain-specific models [3, 8], the two-level modeling paradigm does not specify a predefined set of attributes that can be represented.

The approach presented here is based on exploiting this modeling flexibility of archetypes, which will be used to initially define the data attributes, the context, and the semantics exposed by the data mediation services. In addition, the effort invested over the last decade by the medical informatics community in creating large archetype repositories¹ will be leveraged in the provision of these services.

2.2 Annotation strategies

A significant number of biomedical research projects (see Section 2.3), both in Europe and the United States, have addressed the challenge of semantically integrating heterogeneous data sources. Remarkably, it can be argued that they are all following essentially the same annotation strategy: (1) create a comprehensive domain-ontology, and (2) define detailed mappings (annotations) from each data source element (e.g. attribute) into this ontology.

Both of these steps require a very significant effort. For example, creating a comprehensive ontology, even when reusing parts of existing ontologies, is already a huge undertaking. Similarly, due to the size of the data sources commonly to be integrated, their complete annotation is error-prone and extremely time consuming.

Such an strategy, which must be considered by all accounts as the state of the art, does not scale, and thus can not be used as part of a generic biomedical research infrastructure. In contrast, an incremental approach with lower setup costs, as presented here, is more appropriate.

2.3 Data integration and Mediation approaches

Data mediation and semantic technologies have been subject to intensive research in the e-Health community. For

¹<http://openehr.org/knowledge/>

example, the EU @neurIST project² created a generic IT infrastructure for the management of cerebral aneurysms [3]. Data integration in @neurIST was generic in scope, and was based on the Vienna Grid Environment [4] to mediate different data sources via Web services utilizing OGSA-DAI and OGSA-DQP³. On top of these Web services the project incorporates a domain ontology [6] and rich annotations to setup data mediation services according to selected ontology concepts [14]. Contrarily to our approach, this requires in-depth annotations, which in turn necessitates coordinated efforts of domain, semantic and database experts.

The EU Health-e-Child project⁴ created a modeling methodology based around three complementary concepts: data, metadata, and semantics (similarly to Section 2.1). Ontology-assisted query formulation enabled a clinician to semi-automatically create relational database queries based on ontology concepts [15]. This system required the creation of a domain-specific ontology, and relationships to the underlying databases (annotations). While @neurIST focused on the setup of mediation services based on ontology concepts, Health-e-Child emphasized improving the creation of relational queries based on ontology concepts. However, the huge collaborative efforts required to create an appropriate domain ontology and the corresponding annotations remain a considerable obstacle.

The third related large EU project was ACGT⁵, which developed a semantic Grid infrastructure to support multi-centric clinical trials in cancer research. Alike to the already mentioned projects, the ACGT semantic data integration framework was based on a domain ontology and corresponding mappings to the actual database schemes (annotations), which in turn was also experienced as challenging cross-domain collaboration [8]. Beyond these similarities, ontology-based (semantic) queries in SPARQL are envisaged by the project [15].

The US cancer Biomedical Information Grid (caBIG) project⁶ [16] created an infrastructure for cancer research. Semantic interoperability between federated data sources is achieved with a common object-oriented view on the underlying data-sources, an according ontology and annotations. The US Biomedical Informatics Research Network (BIRN)⁷ [11] employed a mediator architecture to enable access to distributed and heterogeneous data. At its core, the infrastructure also utilizes ontologies and annotations to enable semantic interoperability.

The achievements in all these projects showed that linking ontologies and actual data sources requires significant coordinated effort and has been performed individually for

²<http://www.aneurist.org>

³<http://www.ogsadai.org.uk/>

⁴<http://www.health-e-child.org/>

⁵<http://eu-acgt.org/>

⁶<http://cabig.cancer.gov/>

⁷<http://www.birncommunity.org/>

different domains. As a consequence, reducing these efforts in the future remains a challenge, which is specifically addressed in our approach by leveraging pre-existing work.

2.4 Dataspaces

The rapidly expanding demand for ubiquitous data availability has led the database research community to produce interesting results, but without a central focus or coordinated agenda. The most severe information management challenges faced by organizations stem from the need to exploit a number of heterogeneous and interrelated data sources, but having no means to integrate them in a principled fashion. The concept of *dataspaces* has been proposed [9] as an architectural abstraction of the technical challenges involved in addressing this scenario, which provides a research agenda for data management.

The use-cases of this agenda include, for example, biomedical research groups working on modelling and simulation of complex biophysical processes in the human body [12]. Obviously, many data management and sharing challenges arise in this context.

The essential underlying principle of dataspaces is that a system does not require full integration of data sources in order to offer useful services. In line with this vision, our approach offers an essentially incremental path towards exposing data sources through data mediation services. For example, semantic annotations from standardised ontologies are re-used from existing artefacts (i.e. archetypes), reducing the initial effort required to expose a data source (see Section 3.1). Also, services do not aim to expose complete data sources, but only the concepts that are needed at any given time, following the archetype-based paradigm to information sharing [1]. Additional concepts are exposed in an incremental fashion.

3 Method

This section outlines the core method of our approach, the role played by archetypes in the automatic semantic provisioning of data mediation services, and its advantages as compared to previous work. This description is illustrated with a running example, which clarifies all the steps involved and the challenges to overcome.

It also justifies how, by re-using existing archetype repositories, the potentially daunting task of annotating data sources can be very significantly reduced, leading to an incremental annotation process.

3.1 Associations of ontology concepts with data sources

An essential step during the semantic provisioning of data mediation services includes the *annotation* of the underlying data sources using standard terminologies, like ontologies. These annotations, represented in Fig. 1 as a dotted line, are used to capture the semantics of the data sources, as well as to (semantically) query these sources using standard terms [15].

The semantic enrichment of data sources with these annotations is by itself a significant undertaking, and can become a major bottleneck to the practical applicability of those mediation approaches that are not specifically designed to address this issue [8, 14, 15]. While some automatic annotation mechanisms have been proposed [10], their application in a wide-scale deployment is limited so far because they are unlikely to meet the required level of performance and quality.

The archetype-based mediation advocated in this paper does not require a large initial annotation effort. In contrast, data annotations are defined via archetypes, thus separating the annotations into two independent steps, namely *binding* and *association*, as shown in Figure 1. This approach has several advantages as compared to the 'traditional' annotation process. On the one hand, archetype definitions, by their very nature, commonly include term bindings into well established ontologies (e.g. SNOMED-CT). Thus, this step of the annotation would already be performed by re-using existing archetype repositories (see Section 2.1).

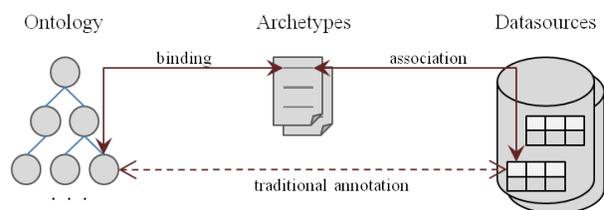


Figure 1. Traditional vs. Archetype-based (bindings plus associations) annotations.

On the other hand, the 'association' step is a simpler process than general annotation, as it is dealing with an individual archetype (which by definition is related only to one single information concept) as well as some (associated) data items in the data source. Due to the close relation between archetype attributes, corresponding data items should also be found in the same or closely (e.g. direct foreign key constraint) related tables in a well structured data source.

Associations can be defined incrementally and on-demand. Once an archetype has been associated with its

related data source fields, it could already become part of a mediated schema. Thus, it already may expose a part of the data source, even if additional archetypes are associated to other fields of the same data source in the future.

Besides requiring less effort to expose data in an incremental fashion, we also expect that existing automatic annotation or pattern matching approaches can be applied to and eventually improve the 'association' step. Presumably, such approaches will lead to better results compared to their application in traditional annotation and standard (large) ontologies (see Section 6).

3.2 Automatic provisioning of data services

In service oriented architectures (SOA) data sources are exposed by dedicated data services to provide a uniform interface to access different kinds of information. Our method also facilitates the SOA-paradigm using the Vienna Grid Environment (VGE) [4] to provision customized data services, which provide access to different data sources in a mediated fashion. These services are usually referred to as data mediation services. It should be noted that the proposed method is generic enough that it could also be applied with a different data service infrastructure, which supports mediation capabilities of different data sources (e.g. SAP's Data Federator⁸), but we have chosen VGE as our target environment as it has been successfully utilized in EC-funded projects like @neurIST or Admire⁹.

In general, the provisioning process of a data mediation service involves (1) the manual creation of the virtual global schema and (2) its mapping to the individual data sources and their respective table columns. Both tasks require substantial efforts if applied in a broader context like the @neurIST project [14] and, thus, an automatic generation of the global schema and the mapping would significantly ease the provisioning of data mediation services. Moreover, the level of abstraction from the users' point of view improves tremendously by utilizing semantic technologies. Obviously, selecting well-established ontology concepts to create a virtual data mediation schema is much more convenient than defining a relational database schema. Besides this primary use-case, our proposed approach also alleviates maintenance of services, because this flexible provision allows dynamic adaptations of the services according to evolving requirements or database structures.

The overall generation of the global schema and the mapping is depicted in Figure 2. The process is based on the bindings and associations of archetype attributes with ontology concepts and data source columns, respectively. By using the bindings information the global schema (i.e. table

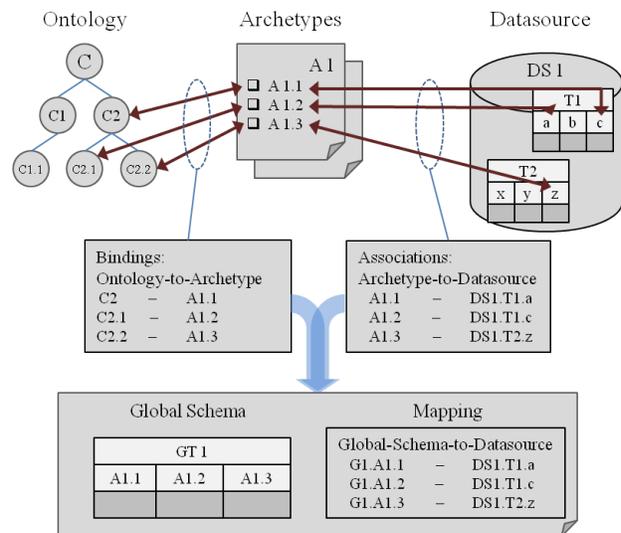


Figure 2. Generating the global schema and mapping from Archetype-based annotations.

GT1) is determined, with the archetype attributes as individual columns of this table, e.g. A1.1. For the sake of simplicity the illustration in Figure 2 shows only a single table for an archetype, but the generation process is generic to support multiple hierarchical tables, if the archetype has such a structure. For the mapping of the global schema to the actual data source the associations of the archetype attributes to the data source table-columns are utilized. In the example of Figure 2 each column of the global table GT1, which is derived from an archetype attribute, is mapped to specific data source, a corresponding table and column (e.g. DS1.T1.a).

The underlying assumption of this automatic generation is that archetype attributes are associated with data source columns which are related items, in the sense that their corresponding tables can be joined. More specifically, two data items in a global table have a common identifier or at least can be joined following the links of their foreign key relationships to a common identifier.

Figure 3 sketches an excerpt of the global schema and its mapping information to individual data sources. In this example the global table GT1 is created by joining the results of two select operations on different source tables. For the sake of simplicity in this case we assume that both source tables (T1 and T2) can be joined using the key id. If an archetype is associated with multiple data sources (for the same attribute), the results of the queries in these data sources are consolidated using the UNION operation of the VGE system.

⁸<http://www.sap.com/>

⁹<http://www.admire-project.eu/>

```

<?xml version="1.0" encoding="UTF-8"?>
<VDSConfig xmlns="...">
  <databaseSchema>
    <logicalSchema>
      <table name="GT1">
        <column fullName="A1 1" >
        <column fullName="A1 2" >
        <column fullName="A1 3" >
      </table>
    </logicalSchema>
    </databaseSchema>
  </MappingSchema>
  <VDSTable table_name="GT1">
    <join result_name="GT1u1">
      <select result_name="GT1s1" dataResource-Ref="URI-TO-DS1.T1." >
        <mapSource>
          <column column-ref="id">
            <source-ref id="source-ref">
          </column>
        </mapSource>
        <column column-ref="A1 1">
          <source-ref id="source-ref">
        </column>
      </select>
    </join>
  </VDSTable>
</MappingSchema>
</VDSConfig>

```

Figure 3. Sample global schema and mapping information.

4 Use cases

This section illustrates the method described in Section 3 with two real use case scenarios motivated by the experiences in recent international biomedical research projects, namely @neurIST and Ricordo.

Firstly, it briefly describes an example archetype, which will be used to guide the provisioning of the semantic mediation services. This archetype is one of the 280 archetypes currently available from the Clinical Knowledge Manager¹⁰, a freely available archetype repository. Then, it details how the archetype and each of the data sources are processed in order to provision the services.

4.1 Archetype example: Blood pressure

The archetype example used in these use cases defines the information required in order to communicate a blood pressure measurement. The actual data values represented in this archetype are shown in Figure 4, namely systolic, diastolic, mean arterial pressure, and pulse pressure. It must be noted that pulse pressure is a derived data value, resulting from the difference between the systolic and the diastolic blood pressure values. It is also of paramount importance to identify the term binding information in which, for example, the archetype term *systolic* is bound to identifier 163030003 in the SNOMED-CT ontology.

In addition to the actual data fields needed for the measurement, archetypes can include further information to fully interpret the medical context. In the blood pressure archetype, this would include a description of the position of the subject at the time of measurement and the body location where this measurement was taken.

¹⁰<http://openehr.org/knowledge/>

Archetype: Blood Pressure (openEHR-EHR-OBSERVATION.blood_pressure.v1)				
Header	Data	State	Protocol	Events
Structure: Tree				
Occurrences: 1..1 (mandatory) Cardinality: 0..* (optional, repeating, unordered)				
Systolic	Quantity Occurrences: 0..1 (optional) [SNOMED-CT(2003):163030003] (On examination - Systolic BP reading (Finding))	Peak systolic arterial blood pressure - measured in systolic or contraction phase of the heart cycle.	Property: Pressure Units: • 0.0..<1000.0 mm[Hg] Limit decimal places: 0	
Diastolic	Quantity Occurrences: 0..1 (optional) [SNOMED-CT(2003):163031004] (On examination - Diastolic blood pressure reading (Finding))	Minimum systolic arterial blood pressure - measured in the diastolic or relaxation phase of the heart cycle.	Property: Pressure Units: • 0.0..<1000.0 mm[Hg] Limit decimal places: 0	
Mean Arterial Pressure	Quantity Occurrences: 0..1 (optional)	The average arterial pressure that occurs over the entire course of the heart contraction and relaxation cycle.	Property: Pressure Units: • 0.0..<1000.0 mm[Hg] Limit decimal places: 0	
Pulse Pressure	Quantity Occurrences: 0..1 (optional)	The difference between the systolic and diastolic pressure.	Property: Pressure Units: • 0.0..<1000.0 mm[Hg] Limit decimal places: 0	

Figure 4. Data fields in the blood pressure archetype.

4.2 @neurIST

The @neurIST project developed a uniform data representation for all clinical information available in the project following a Clinical Reference Information Model (CRIM) [13]. The population with data resulted in a large patient information database distributed across Europe and comprising data of over 1.300 subjects with aneurysms. This database was not materialized in a centralized database, but it was presented to the user as a single (virtual) data source through the mediation services used in the project [3].

The CRIM-schema is rather large, with 91 tables and 1286 attributes. The related blood pressure attributes of the archetype (i.e. systolic, diastolic and mean arterial pressure), can be found quickly in table *EventVitalSigns* with column names *systoPress*, *diastoPress* and *mBloodPress*.

Generating data mediation services for the @neurIST virtual database based on archetypes can be realized seamlessly and in an incremental fashion. This confirms our initial hypothesis, that using our approach dramatically reduces the annotation effort and immediately results in semantically exposed data, even if just a few archetypes are associated with actual data.

4.3 Ricordo

RICORDO¹¹ is a EC-funded project researching a communal ontology-based annotation strategy that supports the interoperability of physiopathology data at multiple biological scales. One of the data sources to be integrated includes clinical information about studies on heart failure. This is a small database of 20 tables and about 300 attributes. As part of routine clinical assessment information, blood pressure measurements are stored in this database, in one single (string) attribute, in the form "systolic/diastolic". This is an

¹¹<http://www.ricordo.eu/>

example of a non-normalized database, commonly found in practice, but this can easily be solved by creating a view which exposes both values independently. The terms from the blood pressure archetype can then be associated with the attributes selected in this view. It must be noted that the term 'mean arterial value', required according to the archetype of Section 4.2, is not recorded in this database. This fact exemplifies an additional benefit of our approach, which can be used to quantify the quality of the information stored in a data source. This would be measured as the number of missing fields according to the associated archetypes.

5 Conclusions

Our approach improves the process of creating semantic mediation services. It re-uses well-established contributions (archetypes) and the VGE, which significantly reduce the burden of annotation and setup of mediation data services. Another benefit of our approach is its incremental nature, as one single association between archetype and data source will already provide a useful service by itself. Additional associations will enrich the services available. Semantic annotations can be added (rather, re-used through archetypes) which will automatically enrich the data sources being exposed.

6 Future Work

This work is now being extended in three main directions. Firstly, the context provided by archetypes facilitates the association of archetype attributes with related data source attributes (as could be done to bind archetype terms [19]). It is expected that most of this association could be done automatically once a single archetype attribute has been associated, through pattern matching techniques. The quality of the resulting automated process would need to be evaluated.

Secondly, it is necessary to provide an appropriate interface for non-technical users (researchers), when querying the mediated global schema. Queries should be build using ontological concepts, which requires semantic resolution, to translate a query into the underlying data sources.

Finally, the definition, setup and maintenance of data mediation services needs to be evaluated in the presence of evolutionary changes to the underlying data sources.

References

- [1] T. Beale. Archetypes: Constraint-based domain models for future-proof information systems. In *Proc. 11th OOPSLA Workshop on Behavioral Semantics*, pages 16–32, 2002.
- [2] G. Bell, T. Hey, and A. Szalay. Beyond the data deluge. *Science*, pages 1297–1298, 2009.
- [3] S. Benkner and A. Arbona. @neurIST - infrastructure for advanced disease management through integration of heterogeneous data, computing, and complex processing services. *IEEE Trans. Inform. Techn. Biomedicine*, 2010.
- [4] S. Benkner, G. Engelbrecht, M. Köhler, and A. Wöhrer. *Virtualizing Scientific Applications and Data Sources as Grid Services*, chapter Cyberinfrastructure Technologies and Applications, pages 81–111. Nova Science Publishers, 2008.
- [5] J. Bisbal and D. Berry. An analysis framework for electronic health record systems. *Meth. Inform. in Medicine*, 2010.
- [6] M. Boeker, H. Stenzhorn, and K. Kumpf. The @neurIST ontology of intracranial aneurysms: Providing terminological services for an integrated it infrastructure. In *Proc. AMIA Annual Symposium*, 2007.
- [7] A. Branson, T. Hauer, and R. McClatchey. A data model for integrating heterogeneous medical data in the Health-e-Child project. In *HealthGrid Conf.*, 2008.
- [8] M. Brochhausen and A. Spear. The ACGT Master Ontology and its applications - towards an ontology-driven cancer research and management system. *Journal Biomedical Informatics*, 2010.
- [9] A. Das Sarma, X. Dong, and A. Halevy. Bootstrapping pay-as-you-go data integration systems. In *Proc. ACM SIGMOD*, pages 861–874, 2008.
- [10] H. H. Do and E. Rahm. Matching large schemas: Approaches and evaluation. *Information Systems*, pages 857–885, 2010.
- [11] J. Grethe and E. Ross. Mediator infrastructure for information integration and semantic data integration environment for biomedical research. *Methods in Molecular Biology*, pages 33 – 53, 2009.
- [12] P. Hunter. A vision and strategy for the virtual physiological human in 2010 and beyond. *Phil. Trans. Royal Society A.*, pages 2595–2614, 2010.
- [13] J. Iavindrasana and A. Depeursinge. Design of a decentralized reusable research database architecture to support data acquisition in large research projects. *Studies in health technology and informatics.*, pages 325–329, 2007.
- [14] K. Kumpf, A. Wöhrer, S. Benkner, G. Engelbrecht, and J. Fingberg. *Grid Computing for Bioinformatics and Computational Biology*, chapter A Semantic Mediation Architecture for a Clinical Data Grid, pages 267–298. John Wiley and Sons, 2008.
- [15] K. Munir, M. Odeh, , and R. McClatchey. Managing the mappings between domain ontologies and database schemas when formulating relational queries. In *Proc. IDEAS*, pages 131–141, 2009.
- [16] S. Oster and S. Langella. caGrid 1.0: An enterprise grid infrastructure for biomedical research. *Journal of the American Medical Informatics Association*, pages 138–149, 2008.
- [17] H. Stenzhol, S. Schulz, M. Boeker, and B. Smith. Adapting clinical ontologies in real-world environments. *Journal of Universal Computer Science*, pages 3767–3780, 2008.
- [18] A. Wöhrer, P. Brezany, and T. A. Novel mediator architectures for grid information systems. *Future Generation Computer Systems*, pages 107–114, 2005.
- [19] S. Yu, D. Berry, and J. Bisbal. An investigation of potential terminological links to archetypes in an external clinical terminology through the construction of terminological Shadows. In *Proc. IADIS Int. Conf. on e-Health*, 2010.