Using the OM^2R Meta-Data Model for Ontology Mapping Reuse for the Ontology Alignment Challenge – a Case Study

Hendrik Thomas, Rob Brennan, Declan O'Sullivan

Federated Autonomic Management of End-to-end Communication Services (FAME),
Knowledge & Data Engineering Group, School of Computer Science and Statistics,
Trinity College Dublin, Ireland
{thomash, rob.brennan, declan.osullivan}@cs.tcd.ie

Abstract. Ontology matching and mapping is of critical importance to effective consumption of distributed and heterogeneous data-sets in today's Web of Data. Since 2004 the Ontology Alignment Evaluation Initiative (OAEI) provides a number of complex challenges to evaluate the performance of the increasing number of matching tools and methods. This leads to the question how the individual OAEI challenges and the individual alignment results can be documented best for effective online consumption, management and further analysis. In this paper, we argue that the current documentation of alignment creation lifecycle aspects within OAEI would benefit from more formal model support. In this paper we present a case study to show how our ontology-based meta-data model for ontology mapping reuse (OM^2R) can be applied for the OAEI to document alignment challenges and some quantification on the likely benefits in terms of helping challenge administrators and participants create consistent documentation in terms of high correctness and less inconsistent statements as well as results that are explicit, predictable and easy to interpret.

Keywords: Ontology Matching, Ontology Alignment, Meta-Data Model

1 Introduction

Ontology matching and mapping is of critical importance to effective consumption of distributed and heterogeneous data-sets in today's Web of Data [1,2]. To support the need for integration the number of methods that are being proposed for matching of ontologies/datasets has increased considerably, which consequently has created the need to establish a consensus for evaluation of these methods [2]. The Ontology Alignment Evaluation Initiative (OAEI) [3] organizes annual evaluation campaigns with the aim of "assessing strengths and weaknesses of alignment/matching systems; comparing performance of techniques; increase communication among algorithm developers" [4]. Each alignment challenge provides a collection of ontologies and reference alignments which enables a comprehensive evaluation of matching tools and their outputs in a controlled environment. In 2012 the OAEI provided seven distinct challenges and each challenge contains up to 58 individual alignment tasks. These challenges and reference ontologies are subject to changes from year to year to provide an even more effective and revealing test bed [3,5]. In the light of the OAEI's goals this leads to the question of how the individual OAEI challenges and the individual alignment results of the participants can be best documented for effective online consumption, management and analysis over time. In other words, for third parties to interpret and evaluate the alignment results of a particular matching method correctly they often need to know precisely how each challenge was conducted. Also any changes to the challenge setups or target ontologies need to be documented clearly as the evaluation needs to be run over several years in order to allow for adequate measurements of the evolution of the field [3].

This creates the need for suitable documentation which can support participating users and researchers in evaluation of the alignment results [2,6]. The standards for such documentation tend to emerge over time as needs are identified and addressed. Since 2004 the OAEI has specified that each challenge must be documented on a specific web page to provide the scaffolding for the participants [4], e.g. including a short textual description of the dataset and evaluation modalities.[1] The majority of this information is provided in text form, lists and some embedded meta-data in the ontologies themselves. We argue that a more formal and structured model for the alignment lifecycle and appropriate alignment management meta-data may have benefits for both organisers and participants including the creation of more consistent documentation and the potential for automated re-use of alignments for other purposes in the future [2,7]. As each challenge is maintained by an independent group such a model can also be of benefit for the OAEI organisers to manage changes to reference alignments and to track submissions over the years to identify performance improvements and trends, e.g. to determine what alignment approaches are becoming more popular and more successful [3]. We argue that an improved meta-data model can help to leverage the experience gained in the OAEI to extend its focus from a pure test platform [8] to a large scale alignments repository [4] which can demonstrate how alignments can be managed, shared and reused over time successfully. To achieve such a shared understanding of matching challenges and the alignment creation in the true sense of the Semantic Web [9] a meta-data model needs to be documented clearly to help users understand the intended meaning of the individual fields easily [10,11]. To support analysis and reuse it needs to be formally detailed in a machine-interpretable notation such as OWL. It must promote the creation of consistent documentation instances in terms of correctness and avoidance of inconsistent statements.

In parallel to the work of OAEI, the authors have developed an ontology-based meta-data model for ontology mapping reuse (OM^{2}R) [7,12,13]. Thus OM^{2}R has a broader scope of supporting ontology mapping (alignment) management. Nonetheless at least part of the OAEI activity can be viewed as a very large-scale alignment management exercise, especially with respect to the historical result-sets. The challenge addressed in this paper is thus: can OM^{2}R be usefully applied to supporting OAEI activities and some quantification on the likely benefits in terms of helping challenge administrators and participants create consistent documentation in terms of high correctness and less inconsistent statements, experimental results that are explicit, predictable and easy to interpret. The model can also support matching retrieval and reasoning about matchings.

In this paper we present a case study to evaluate how the OM^{2}R model can be applied for the OAEI competition to document the alignment challenges to support machine-based online consumption, processing and further analysis of the submitted results through the publication of annotated OAEI challenges, data-sets and result-sets as linked data using the OM^{2}R vocabulary. In this first case study we have selected the benchmark dataset as a representative challenge from the OAEI initiative 2012 [4] and we will evaluate the individual meta-data fields proposed in OM^{2}R in relation to the current documentation.

[1] Please find more details on http://oaei.ontologymatching.org/doc/oaei-submitting.1.html

Please note the OM^2R was designed with a focus on ontology mappings but OAEI focuses on ontology alignment or matching [7,13]. In our terminology matchings are machine-generated correspondence candidates, an essential step in the creation of mappings which are confirmed correspondences created in the mapping phase as part of the overall ontology mapping creation lifecycle [14].

This paper is structured as follows. Section two gives a brief overview of other related meta-data models for ontology matchings. In section three we will provide a brief introduction to the OM^2R model. In section four we will discuss how OM^2R can be applied for the benefit of the OAEI initiative. The paper concludes with a summary and an outlook.

2 Related research

The need for a suitable meta-data model to document ontology matchings has been recognized in the current literature. For example J. Euzenat stated that one of the ten major challenges for ontology alignment is that management "must be complemented with rich metadata allowing users and systems to select the adequate alignments based on various criteria." [2,6] J. Euzenat and his team addressed this need by creating the ontology alignment format which offers a matching representation and basic meta-data identifying the addressed ontologies. Also an extended vocabulary [15] allows some meta-data to be embedded within the format.[2] In addition, EDOAL an expressive and declarative ontology alignment language extends the alignment format [22]. It provides a more detailed documentation of the matching algorithm elements but similar to the ontology alignment format it does not focus on the actual mapping creation lifecycle and management aspects.

Furthermore, we acknowledge the work of other authors in this area [16, 17]. For example N. Noy et al. proposed a community-driven ontology matching tool for public alignment reuse. This system annotates mapping elements in a given format but does not address the creation lifecycle or mapping reuse.

In addition, our work needs to be placed in context with ontology meta-data initiatives like the OMV (a meta-data model for ontologies and related entities [18]) or the PROV-DM (W3C data model for provenance interchange) [19]. These vocabularies can be used to express specific aspects of mappings efficiently like provenance, availability and statistics. Also important is the growing application of matchings in the linked data community to improve the interoperability between these still only loosely coupled data sets [16, 20]. The effort to distribute the matching creation tasks between different parties is increasing which implies the need for users to be able to assess the quality of matching and assess a possible reuse [4].

The current challenge for alignments management and therefore for the OAEI can be summarized as a need for a "convenient and interoperable support, on which tools […], can rely in order to store and share alignments. This involves using standard ways to communicate alignments and retrieve them. Hence, alignment metadata and annotations should be properly taken into account."[2].

The above discussed meta-data models demonstrate how other researchers have addressed these issues but their approaches are limited in the light of the OAEI documentation requirements as they are either focused on the representation of alignment correspondences and not on creation and management related meta-data data or the models are not specific and detailed enough for the alignment management

2 More information can be found on: http://alignapi.gforge.inria.fr/labels.html

and reuse. The OM^{2}R can benefit from their contributions but we argue that the wider objective OM^{2}R which focuses on the whole ontology matching and mapping creation lifecycle can better support the creation of documentation to support retrieval, management and analysis over time for the OAEI.

3 Overview of OM^{2}R

3.1 Basic principles

The main design objective of OM^{2}R was to create a meta-data model for ontology mappings which covers the complete lifecycle including the matching phase to support mapping discovery and management [7,12]. Various formats are available to document ontology matching and mappings [12]. The design of a mapping representation which fulfils all possible requirements for expressing the correspondences might be overly complex, hard to enforce consistency on or alternatively represent only the lowest common denominator information [2, 12]. In contrast, a meta-data layer which documents the mapping lifecycle can complement existing mapping representations. Thus OM^{2}R is used to provide a common vocabulary for documenting mappings but is kept distinct from the mappings themselves. Hence OM^{2}R does not replace existing mapping representation languages but it compliments them with extensive lifecycle and context information which references the actual alignment themselves in a language neutral way. OM^{2}R meta-data is intended to be shared between users and applied in different contexts. Thus unambiguous meaning in terms of a shared common understanding of the documentation fields is essential. Hence OM^{2}R is expressed in an ontology which describes the meta-data structures and embeds extensive descriptions of the model elements (e.g. a short name, a definition, acronyms and a unique identifier) inside the actual model. The ontology contains 38 classes and 21 typed object relations between the individual meta-data fields which can be interpreted by editors, e.g. to enable highlighting of compatible field options. OWL-DL was used to model OM^{2}R instead of RDF(S) because it provides the necessary expressivity and supports greater reasoning to reveal implicit knowledge [7]. In our view, the key to understanding how a particular mapping was created lies in the ontology mapping lifecycle. In other words, the individual phases of the life cycle are used as the basis for the structure of the OM^{2}R and the involved activities provided an indication of what aspects need to be documented in meta-data fields. A common agreement on the phases involved in a full ontology mapping lifecycle has not yet emerged [7,14]. Please find below a mapping lifecycle proposal based on [14] which was used for OM^{2}R:

1.) **Characterisation phase:** The focus of this phase is the discovery of the ontologies which are subject of the mappings in term of the identification of the ontologies and their nature with respect to their amenability for matching methods.
2.) **Matching phase**: The objective of this phase is the description of identification of mapping candidates, either identified by manual selection or by automated matching algorithms [9,20].
3.) **Mapping phase:** The third stage involves the generation of information necessary for the execution of mappings as well as the creation of confirmed mappings.
4.) **Execution phase:** The identified committed and approved mappings can then be rendered into different mapping formats in order to enable processing and sharing.
5.) **Management phase:** Ontology mappings generated in the previous phases need to be managed and maintained until their withdrawal. This includes the sharing of mapping information with third parties, the integration of mapping into other mapping applications.
6.) **Meta-Data creation:** Conceptually a parallel activity to the phases above where meta-data is collected and processed, e.g. automatically extracted from ontologies or manually entered by involved stakeholders. Appropriate tool support may integrate it into the other lifecycle phases.

The key contribution of the formal OM^{2}R model is that it can support the creation of consistent documentation that is suitable for automated consumption and processing. More specifically our model can contribute to the following consistency aspects [21]: structural consistency, logical consistency and application consistency. Each is described in more detail below.

Structural Consistency ensures that the ontology obeys the constraints of the ontology language with respect to how the constructs of the ontology language are used [21]. The OM^{2}R model provides a common set of concepts and relations, thus a clear documented template allowing two users to express their facts by using the same vocabulary and semantic.

Logical consistency sees the ontology as a logical theory, which considers an ontology as logically consistent if it does not contain contradicting information [21]. By explicitly modelling allowed and appropriate relationships, the OM^{2}R model contains information about compatible relations between meta-data fields. For example if an ontology was expressed in the notation RDF/XML and in the formal language RDF(S), this reflect a compatible relation between the notation and formal language used which is modelled explicitly in the OM^{2}R. Our mapping documentation tool based on OM^{2}R can use these relations to highlight logical consistent options in the UI to support the editing process.

Application consistency relates to aspects not captured by the underlying ontology language itself, but rather given by some application or usage context [21]. In our context this relates to the ability of OM^{2}R to support the actual correctness of documentation in relation to a given matching and mapping management scenario.

The actual OM^{2}R model is available for download.[3] Please note beside the OWL file we provide on the same page the Protegé project files which enables you to start using the model to document your own matchings straightforward.

2.2 Evaluation

To validate the OM^{2}R we conducted a wide-scale end-user evaluation experiment with 50 participants drawn from the semantic web research community in 2010. The hypothesis was that the proposed OM^{2}R fields and their structure are considered relevant by users for a mapping reuse decision. The participants were given two mapping documentation scenarios and could rate the relevance of the individual fields for documentation and a reuse decision. The data showed that information identifying the addressed ontologies and matchings (e.g. names and location) are considered most relevant closely followed by details about the specific matching and mapping process used. Overall all of the 29 meta-data fields were considered relevant[4].

In 2012 we conducted a more practical task-oriented experiment with the hypothesis that OM^{2}R can support the creation of consistent documentation (see section 2.2) of the ontology mapping lifecycle and is usable by novice and experienced users in ontology mappings. The users were presented with an editing interface based on the OM^{2}R and asked to document the identification and matching phase of a sample matching scenario based on textual instructions. We used precision and recall [10] as an indicator for the level of achieved application and logical consistency. Overall 48 users completed the experiment with a ration of 40% experts with previous matching experience and 60% novice users with no experience. The following table shows the data we collected:

[3] The OM^{2}R model can be downloaded from: http://www.modelmapping.org/om2r

[4] The % of users who rated a field as relevant ranged from 77% to 23% with a mean of 60%

Metric	All Participants	Expert users	Novice
Application – recall	78 %	78.5 %	77.6%
Application – precision	81.8 %	79.1%	83.6 %
Logical- recall	86 %	91 %	82.2 %
Logical – precision	85 %	85.8%	84.6 %

Tab. 1 Average metrics for application and logic consistency

This evidence supports our claim that the OM^{2}R can support users in the creation of consistent documentation. Also we could not find any statistically significant difference between the support for experts and novice users.

4 Application of OM^{2}R to the OAEI

In this section we discuss the current documentation provided by the OAEI and show how the OM^{2}R can help to add an additional beneficial documentation layer.

4.1 General Approach

To show the benefits of the OM^{2}R an understanding of the involved stakeholders is needed. Please find below an overview:

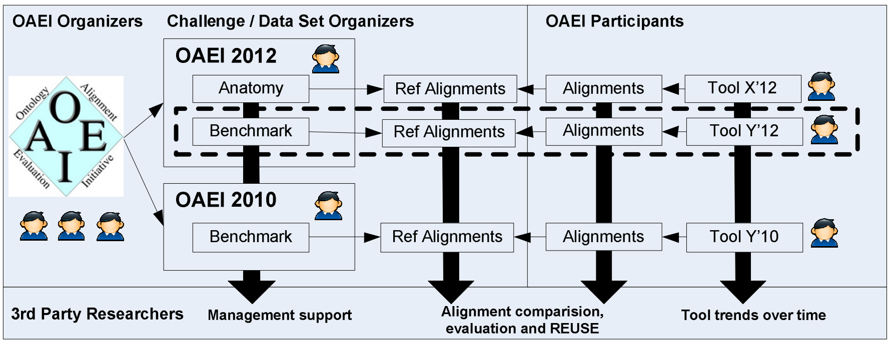

Fig. 1 Overview of the OAEI Stakeholders

The first involved group are the OAEI organizers which are responsible for the overall management, the submissions and the publication of the results for each OAEI initiative per year. Each individual challenge is maintained by an independent group who manages the different alignment tasks, ontologies and reference alignments. Also involved are the actual participants who use their matching tools to complete the individual tasks by submitting alignments or since 2011 their applications as a bundle. The fourth stakeholders are 3rd party researchers, who utilize the results published by the OAEI committee to learn more about the performance of the matching tools based on a metric approach [2]. We argue that an analysis of the reference ontologies, the actual alignments created by the participants as well as the provided reference alignments are of similar interest and value.

The current documentation provided by the OAEI is focused on individual challenges and the different initiatives per year. Each challenge is documented on a specific web page. This web page represents the main documentation source and provides the participants with the needed information to join the challenge. The primary focus is on online consumption as the majority of information is presented in text form, tables and some few meta-data fields embedded in the reference ontologies and alignments. The dotted line in figure 1 indicates the addressed stakeholder of this horizontal documentation focus.

We argue that the OM^{2}R can provide an additional meta-data layer which can extend the current documentation with a more formal model to address the particular needs of 3rd party researchers and organizers. OM^{2}R allows users to create more consistent (see section 3.2), easier to interpret and more explicit documentation which can help to identify trends easier as well as an enable a more detailed comparison of the results of individual contributors over time. We argue that the OM^{2}R can bring the current available information together, add more structure combined with a higher level of detail and a time dimension (big black arrows in fig 1). This can help OAEI organizers and 3rd party researcher to keep a better overview and to manage changes of data sets over time.

To achieve this objective, the OM^{2}R uses a different representation approach for meta-data. Instead of a focus on text designed for human consumption it focused on retrieval and automated processing. It targets specifically the objects of interest for matching embedded in a lifecycle structure. The OM^{2}R is expressed as an ontology and therefore all meta-data information are stored as explicit and meaningful triples, e.g. om2:source_ontology hasNotaton rdf/xml = object of interest - typed relation - meta-data field option. Also explicit relations between the field options are included in the model, e.g. compatible relation between language and notation. This rich index structure makes the editing and the interpretation of the intended meaning easier, less ambiguous and provides a better structure for human and automated consumption. Also the current documentation is limited to single data sets per initiate. This is well suited for challenge participants but limits the view for researchers and organizers. The benefit of the OM^{2}R is that multiple alignments can be documented in one OM^{2}R model. This is particular relevant for the benchmark data set which is designed to be stable over time but as the web page points out the reference ontology has changed in 2010. Comparison, retrieval and reasoning can be supported better if the reference ontologies and their individual alignment versions per year could be documented in one OM^{2}R.

4.2 Meta-Data overview

To gain a more detailed understanding of the individual meta-data that is typically provided in OAEI we will focus in the following sections on one representative alignment challenge. More specifically we demonstrate the contribution of the OM^{2}R for the OAEI by focusing on the meta-data provided for the characterisation and matching phase of the lifecycle.

The selected challenge needs to be extensive in order to provide sufficient context for documentation and was used in previous OAEI initiative in order to allow a comparison of the available meta-data over time. In the latest OAEI challenge in 2012 the following data sets were provided: Benchmark, Anatomy, Conference, Multifarm, Library, Large Biomedical Ontologies, Instance matching [4]. If we consider the last four OAEI challenges (year 2012, 2011.5, 2011, 2010), only the following data sets have been used in all four challenges: Benchmark, Anatomy, Conference. If we compare the provided documentation for the 2012 challenge we can see, that the documentation webpage for the benchmark data set contains the most detailed documentation and therefore the highest amount of meta-data information.[5] It can therefore provide the most insight and will be the focus of our discussion.[6]

---

[5] The word count for the benchmark page was 3505, for anatomy 702 and for conference 544.

[6] Please see for details: http://oaei.ontologymatching.org/2012/benchmarks/index.html.

The following table provides an overview of the individual meta-data fields which have been rated by our end user experiments as relevant (see section 3.2) for the identification, characterisation and matching phase. It shows which information are provided by the OAEI and the corresponding fields in the OM^{2}R. Please note the column "OAEI Fields" indicates if the meta-data information is presented by the OAEI in an explicit field (e.g. embedded in the ontology) or was mentioned in an unstructured text segment. The column also tells you if the information is available for all (A) addressed target and source ontologies or only for some (S). Following the table the individual lifecycle phases are discussed in more detail:

Meta-Data field	OAEI Fields	OM^{2}R - Meta-Data Fields
Name of ontologies	Text (A) Field (S)	SourceOntology :Om2r:human_readable_name: "Biology Top Level Ontology"
Description of ontologies	Text (S) Field (S)	Om2r:description
Location of ontology	Text (A)	Om2r:hasLocation (type url)
Creation date of ontologies	Field (S)	Om2r:hasCreationDate (type date)
Unique identifier for ontologies	Field (A)	Om2r:hasIdentifier
Ontology Version	Missing	Om2r:hasVersion (URI)
Complexity of the ontology	Text (S)	Om2r:hasClassCount 73, hasInstanceCount 3 hasPropertyClass 3
Design of the ontologies	Text (S)	Om2r:hasDesign om2r:deep_hierarchy.
Notation of Ontologies	Text (S)	Om2r:hasNotation RDF/XML
Formal Language of Ontologies	Text (S)	Om2r:hasFormalLangauge OWL
Matching Location	Text (A)	Matching Om2r:hasLocation: www (URL)
Formal Language of the Matching	Test (S)	Om2r:hasformalMatchingLanguage: EDOAL
Notation of the Matching	Missing	Om2: hasNotation: RDF/XML.
Matching Method	Missing	Om2r:hasMethod (manual, automatic, mixed)
Matching Tool	Missing	Om2r:isTool AlignmentServer
Matching Algorithm	Missing	Algorithm :encodedIn: Java, Algorithm :hasJavaClass: org.stringComp, Algorithm :hasSource: freecode.org/a.zip
Algorithm is based on	Missing	Om2r:isBasedOn rdfs:label, rdfs:class
Applied Threshold	Missing	Om2r:has_Applied_Threshold
Matching Scope	Missing	Om2r:hasScope (complete or partial)
Matching Requirements	Missing	Om2r:hasMatchRequirements (text)

Tab. 2 Comparison of OM^{2}R meta-data fields with the OAEI

In the following sections we will discuss the provided meta-data in more detail. However space is limited here and it recommended that readers download the OM^{2}R ontology for themselves (see footnote 3) and use their preferred tool to explore it.

4.3 Phase 1.1 - Identification of the addressed ontologies

To begin a challenge a participant requires details about the addressed ontologies. On the web page of the benchmark data set a brief description is provided for the source ontology which is referred to as "reference ontology" but also as "bibliographic ontology" and in the task section as "test". Furthermore the web page lists 58 specific tasks where the target ontology is specified, for example [4]:

104) Concept test: Language restriction – This test compares the ontology with its restriction in OWL Lite (where unavailable constraints … Ontology : [RDF/XML] [HTML] Alignment : [RDF/XML] 201[-2-4-6-8]) Systematic: No names - Each label or identifier is replaced by a random one. Ontology : [RDF/XML] [HTML] Alignment : [RDF/XML] [HTML]

Please note the amount of descriptive information for the target ontologies is not consistent for each task, e.g. see example for test 104 vs. 201. Please note that the tasks listed on the lower part of the page contain less information than on the top. Some of the target ontologies have additional meta-data embedded in their source code, e.g. <dc:description>, <rdfs:label> for task 225 but these information can not be

found consistently, e.g. are missing for task 250 and 303. The provided alternative names and descriptions are quite suitable for participants. However, a more consistent approach is needed to support retrieval, analysis and automatic processing. Thus the OM^2R provide the following explicit fields: (**Target and Source) Ontology Name** and **(Target and Source) Ontology Description field.** Thanks to the ontology approach additional meta-data can be expressed easily and meaningful, e.g. hasAlternativeName e.g. hasNaturalLanguage “German”.
In addition, the data set provides information where the sources of the addressed ontologies can be downloaded. The OM^2R provides similar information but in an explicit field **(Target and Source) Ontology Location** to allow automated system to retrieve the required information which currently can be difficult, e.g. the source for the reference ontology points to a section on the web page rather to the actually file.

To track down changes of reference ontologies over time or to negotiate a possible reuse of an ontology it is essential to be able to contact the authors. Currently only few contact information are embedded in some of the reference ontologies, e.g. <dc:contributor>Antoine Zimmermann antoine.zimmermann@inrialpes.fr </dc:contributor>). To promote the publication of such information, the OM^2R provides a dedicated field for this purpose: **(Target and Source) Ontology Editor**. Please note, to simplify the population existing ontology templates for contact details can be used in the OM^2R to help identify the creator more accurately, e.g. FOAF: Ontology creator om2r:firstName Hendrik, Ontology creator om2r:surname Thomas.

For an analysis over time information about the current version of the ontologies and their changes are critical. A good indicator is the creation time and the OAEI provides some textual references. As various date formats exist, an explicit and unambiguous representation is helpful to avoid confusion which why the OM^2R provides the field **(Target and Source) Ontology Creation Date** for an explicit time and date of the creation of the ontology. Internally the date will be represented as a set of explicit triples: CreationDate :hasYear: 2010, CreationDate :hasMonth: 5, CreationDate :hasDay: 4, CreationDate :hasTimeZone: MEZ.

Another relevant aspect is specific information about the changes to reference ontologies as they can create a bias for comparison of results over time. For example, in 2012 the web page states that the reference ontology for the benchmark data set has been altered and “The test is not anymore based on the very same dataset that has been used from 2004 to 2010. We are now able to generate undisclosed tests with the same structure. They provide strongly comparable results and allow for testing scalability.” [4] but no further details are provided. The OM^2R can assist in providing a more detailed and structured documentation of changes with:

Ontology Version provides details about the specific version entered by the editor and a simple hashId to enable an automated and unbiased check for differences: om2r:asVersionId and om2r:hasHashID. Also the **Ontology Change Log** fields can contain elements with a short textual description of the specific conducted changes.

For humans names are a dominant key for identification but in the Semantic World an unambiguous identifier for the ontologies is essential to allow automated processing. In the data set the base url of each ontology is used for this purpose which is unique for each challenge and each data set, e.g. <rdf:RDF xml:base="http://oaei.ontologymatching.org/2012/ benchmarks/ 250/onto.rdf#"> for the task 250. Till 2010 the web page claims the same ontology was used for this dataset but each ontology has a unique identifier and is therefore potentially different. To avoid any miss interpretations the OM^2R provides an explicit field **Ontology Identifier** where unique identifier can be stored.

4.4 Phase 1.2 – Characterisation of the addressed ontologies

Information about the language aspects of the ontology files are of crucial importance for processing and compatibility issues of editing tools. The OAEI provides information about notation and the formal language in text form on the data set page, e.g. the web page states that the reference ontology is available in rdf/xml. The formal language is mentioned in the text but not consistently and in some cases missing, e.g. see the example description for task 236 in section 4.3. To help users in interpreting and reusing the provided resources more explicit information can be helpful, e.g. reasoning can only be applied to OWL DL not OWL Lite, thus stating the language as OWL would be too broad. OM^2R addresses this issues with the following fields: **Ontology Formal Language:** An ontology language is a formal language used to encode the ontology. As there are a number of such languages this field specifies the language, :hasFormalLanguage: http://www.w3.org/2002/07/owl. In case of OWL it is important to specify the sublanguage, too e.g. :subLanguage: OWL-DL. **Ontology Notation**: Beside the ontology language, the specific exchange notation used to represent the addressed ontology can be specified which is essential for tool support and exchange, e.g. TargetOntology :hasNotation: RDF/XML.

The next relevant area for 3rd party researchers and participants is the complexity of the addressed ontologies which is an essential factor when choosing an appropriate algorithm. For this purpose the OAEI provides information about the size of the source ontology (e.g. number of classes) but only in text form and not for the target ontologies. To support analysis and to judge the performance results in relation to the complexity more explicit fields can be helpful to allow automatic harvesting and processing. The OM^2R can assist the publication of these information with the following fields for the target and the source ontology:

Ontology Size: An explicit statement of the amount of classes, properties and instance, e.g. om2r:TargetOntology om2r:containsAmountOfClasses: 50

Ontology Design: Provides an indication of the basic design of the ontology, e.g. a sophisticated and deep hierarchy, a flat class hierarchy with few parent-client classes. The motivation for this field is to provide a broad classification, as different matching algorithms are more suitable for certain structures and size information alone are not sufficient enough, e.g. om2r:TargetOntology om2r:hasDesign om2r:flat_hierarchy.

4.5 Matching phase

The next area of relevance which was identified in our studies (see section 3.2) are details about the matching representation. This refers to the provided gold standard per dataset task and the individual submissions of the participants. The OAEI provides a location where the alignment can be downloaded.[7] In the OM^2R we provide the following explicit field for the location: **Matching Ontology Location.** This is a URL where the file can be downloaded, e.g. Matching :hasLocation.

In regards to the language aspect only the description text per task indicates that the alignment is expressed in XML/RDF but no information are provided for the formal language, e.g. EDOAL. In the OM^2R we provide the fields: **Matching Language:** A matching language is a formal language used to encode the correspondences, e.g. :hasFormalMatchingLanguage: om2r:edoal. In addition, the

[7] Please note we observed an inconsistency in regards to provenance and location, as in the 2012 challenge the alignment links in task 104 points to the 2011 challenge http://oaei.ontologymatching.org/2011/benchmarks/104/refalign.rdf

specific exchange **Matching Notation** can be specified which is essential for tool compatibility and reuse, e.g. Matching :hasNotation: RDF/XML.

Another key aspect are details about the actual method used to created the alignments. We can note that for all 58 tests a gold standard reference alignment is provided but most of the representations do not provide any information about the method or tool used to generate them. The alignment format provides a corresponding meta-data field like <method> for information on the applied matching class but none of this information have been provided in OAEI 2012 benchmark data set.

The OM^2R support the population of these information with the following fields. Please note the OM^2R provide specific instances for all fields which a user can select during the editing process and for each field option the compatible options in related fields are documented, e.g. compatible matching tools for matching methods.

Matching Method: Which generic method was used to find suitable candidates for a matching in the addressed ontologies? Om2r:hasMethod – manual, automatic, mixed
Matching Tool: Specified the tool which was used to generate the alignment, .e.g. hasMatchingTool
Matching Algorithm: If an automated selection was applied, this section provides a descriptive, human-readable label to identify the matching algorithm used. For example: matching :basedOn: Levenshtein distance, Levenshtein distance :isDefinedIn: http://en.wikipedia.org/wiki/Levenshtein_distance
Matching Algorithm Implementation: A descriptive, human-readable label to identify the specific implementation of the algorithm. Could be a URL or a specific JAVA class name like org.jena.stringComparsion. Also helpful is to provide a URL to download the source code. For example: Algorithm :encodedIn: Java, Algorithm :hasJavaClass: org.jena.stringComparsion, Algorithm :hasSource: http://www.freecode.org/123.zip
Applied Threshold: Defines the specific value of the similarity measure which needs to be passed in order to justify a matching pair based on the assumptions of the individual algorithm, e.g. om2r:has_Applied_Threshold. More complex methods may need multiple thresholds or iterations to be modeled instead.
Matching Scope: Defines the scope or area the matching is applied. In particular if all elements are matched to each other or only a particular subset, e.g. om2r:hasScope – complete or partial
Element Matching is based on: Defines the elements which are analyzed by the algorithm to identify the matching pairs, e.g. RDFSLabelForClass
Matching Requirements: Provides details of the specific requirements which needed to be fulfilled to apply the matching, e.g. hasMatchRequirements (text)

5 Conclusions and Final Remarks

In this paper we presented a case study to show how our ontology-based meta-data model for ontology mapping reuse (OM^2R) can be used to extend the current documentation of the OAEI for alignment challenges. We showed how the OM^2R can help administrators and participants create more consistent documentation instances in terms of high correctness and less inconsistent statements as well as support 3rd party researchers with more explicit, detailed, predictable and easy to interpret documentations. We argue that an improved meta-data model can help to leverage the experience gained in the OAEI to extend its focus from a pure test platform [8] to a large scale alignments management repository [4] which can demonstrate how alignments can be managed, shared and reused over time successfully. The overall objective of the OM^2R is to support the sharing of a common understanding of the ontology matching creation and application lifecycle which can hopefully provide a positive contribution to promote and support the reuse of alignments outside the current testing scope.

Acknowledgement

This work is partially funded through the Science Foundation Ireland FAME Strategic Research Cluster (award No. 08/SRC/I1408), http://www.fame.ie.

References

1. Marshall S. et al: Emerging practices for mapping and linking life sciences data using RDF - a case series. In: *Journal of Web Semantics: Science, Services and Agents*, Volume 12, 2012.
2. Shvaiko, P., Euzenat, J.: Ontology matching: state of the art and future challenges. In *Journal of IEEE Transactions on Knowledge and Data Engineering*, 2012 in press.
3. Euzenat, J., Meilicke, C., Shvaiko, P. et al.: Ontology Alignment Evaluation Initiative: six years of experience, In: *Journal on Data Semantics,* Volume XV Edition 6720, pp 158-192, 2011.
4. Euzenat, J.: Ontology Alignment Evaluation Initiative, main homepage, 2012 http://oaei.ontologymatching.org/
5. Euzenat J. et all: Final results of the Ontology Alignment Evaluation Initiative 2011, In: *The Sixth International Workshop on Ontology Matching: Proceedings of the 10th International Semantic Web Conference* ISWC-2011, 2011.
6. Shvaiko, P., Euzenat, J.: Ten Challenges for Ontology Matching. In*: Proceedings of the 7th International Conference on Ontologies, DataBases, and Applications of Semantics* (ODBASE), 2008.
7. Thomas, H., Brennan, R. O'Sullivan, D.: MooM - a Prototype Framework for Management of Ontology Mappings, In: *Proceedings of the 25th IEEE International Conference on Advanced Information Networking and Applications*, Singapore, 22-25 March, 2011, IEEE, pp 548 – 555.
8. Euzenat, J. et al: Results of the Ontology Alignment Evaluation Initiative 2011, In: *Proceeding of the 6th ISWC workshop on ontology matching*, pp 85-110, 2011.
9. Gruber, T. R.: Towards Principles of Ontologies Used in Knowledge Sharing. In: *International Journal of Human-Comper Studies*, Nr. 43, pp. 907-928, 1994.
10. Baeza-Yates, R., Ribeiro-Neto, B.: Modern Information Retrieval. New York, NY: ACM Press, Addison-Wesley, pp. 75 – 78, 1999.
11. Fugmann, R.: Subject Analysis and Indexing: Theoretical Foundation and Practical Advice. Frankfurt a. M., Indeks, 1993.
12. Thomas, H., O'Sullivan, D., Brennan, R.: Ontology Mapping Representations: a Pragmatic Evaluation. In: *21st International Conference on Software Engineering and Knowledge Engineering,*. SEKE 2009, 1 - 3 July 3, Boston, 2009, pp. 228 - 232.
13. Feeney, K., Brennan, R., Keeney, J., Thomas, H., Lewis, D., Boran, A., O'Sullivan, D., Enabling Decentralised Management through Federation, In: *Journal Elsevier Computer Networks*, Volume 54, Issue 16, November 2010.
14. O'Sullivan, D., Wade, V., Lewis, D. Understanding as We Roam, In: *IEEE Internet Computing Journal*, Volume 11, Issue 2, 2007, pp. 26 – 33.
15. Euzenat, J.: *Alignment API and server*, (version 3.2) https://gforge.inria.fr/docman/view.php/117/5036/alignapi.pdf, 2008.
16. N. Noy, N. Griffith, and M. Musen. Collecting community-based mappings in an ontology repository. In *Proc. of International Semantic Web Conference (ISWC)*, Karlsruhe, Germany, 2008.
17. Ghazvinian, A., Noy, N. F. , Jonquet, C. et al : What Four Million Mappings Can Tell You About Two Hundred Ontologies, In: *Proceedings of International Semantic Web Conference (ISWC)*, Washington DC 2009.
18. Palma, R. Hartmann J, Hasse P.: Documenation Report*,* http://surfnet.dl. sourceforge.net/project/omv2/OMV%20Documentation/OMV-Reportv2.4.1.pdf, 2009.
19. Moreau, L., Missier, P: PROV-DM: The PROV Data Model, W3C Working Draft, http://www.w3.org/TR/prov-dm/, July 2012.
20. Millard, I., Glaser, H., Salvadores, M. and Shadbolt, N.: Consuming multiple linked data sources: Challenges and Experiences. In: *First International Workshop on Consuming Linked Data (COLD2010)*, Shanghai, 2010.
21. Haase, P., Stojanovic, L.: Consistent Evolution of OWL Ontologies. In proceeding of the 5th European Semantic Web Conference, pp. 182-197, 2005.
22. David, J., Euzenat, J., Scharffe, F., Trojahn dos Santos, C.: The Alignment API 4.0. In: Semantic web journal Volume 2, Nr.1 pp. 3-10, 2011.