A Justification-based Semantic Framework for Representing, Evaluating and Utilizing Terminology Mappings

Sajjad HUSSAIN[a], Hong SUN[b], Gokce B. Laleci ERTURKMEN[c], Mustafa YUKSEL[c], Charles MEAD[d], Alasdair J. G. GRAY[e], Kerstin FORSBERG[f,1]

[a]INSERM, U1142, LIMICS, AP-HP, F-75006, Paris, France
[b]Advanced Clinical Applications Research Group, Agfa HealthCare, Gent, Belgium
[c]Software Research, Development and Consultancy Ltd., Turkey
[d]W3C Health Care and Life Sciences IG
[e]School of Mathematical and Computer Sciences, Heriot-Watt University, UK
[f]AstraZeneca, Sweden

Abstract. Use of medical terminologies and mappings across them are considered to be crucial pre-requisites for achieving interoperable eHealth applications. However, experiences from several research projects have demonstrated that the mappings are not enough. Also the context of the mappings is needed to enable interpretation of the meaning of the mappings. Built upon these experiences, we introduce a semantic framework for representing, evaluating and utilizing terminology mappings together with the context in terms of the justifications for, and the provenance of, the mappings. The framework offers a platform for i) performing various mappings strategies, ii) representing terminology mappings together with their provenance information, and iii) enabling terminology reasoning for inferring both new and erroneous mappings. We present the results of the introduced framework using the SALUS project where we evaluated the quality of both existing and inferred terminology mappings among standard terminologies.

Keywords: Terminology mapping, semantic interoperability, reasoning, validation

1 Introduction

Achieving a computable semantic interoperability (CSI) among different healthcare applications is – at its core – deeply dependent on the use of "controlled terminologies" which enable the inter-machine exchange of clear and computationally unambiguous semantics [1]. Aiming towards this goal, clinical experts go through a process of defining *terminology mappings* between standard terminologies developed by standards organizations (CDISC, IHTSDO, ICH, etc.), as well as local/legacy terminologies to support a number of specific CSI use cases including (but not limited to):

[1] Corresponding Author.

(i) semantic interoperability between clinical care and clinical research systems for pharmacovigilance and patient safety, (ii) clinical decision making, and (iii) clinical data integration and mediation in pharmaceutical R&D.

Recent research projects focusing on CSI—such as SALUS [2], Open PHACTS [3], and EHR4CR [4]—have published their experiences in defining and utilizing multiple mappings between various terminologies [5].

On the surface, it may appear to the uninitiated as a simple exercise like "this term in this terminology is the same as that term in that terminology." However, it is often a considerably challenging task due to:

- Availability of up-to-date information to assess the suitability of a given terminology for a particular use case.
- Difficulty of correctly using complex, rapidly evolving terminologies.
- Differences in granularity between the source and target terminologies.
- Lack of semantic mappings in order to completely and unambiguously define computationally equivalent semantics.
- Lack of provenance information, i.e. how, when and for what purposes the mappings were created.
- Time and effort required to complete and evaluate mappings.

For example, considering SNOMED-CT as a hub terminology, both ICD-9-CM and MedDRA codes are mapped to SNOMED-CT codes, e.g. "Anaphylactic shock due to serum" (SNOMED.CT) has been (manually) mapped to one IC9-9-CM code and one MedDRA code (see Figure 1).

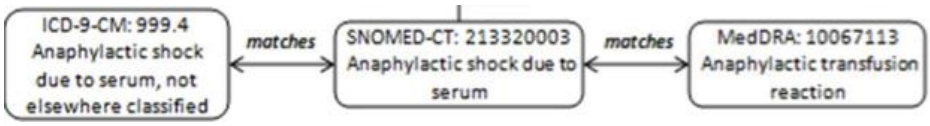

Figure.1. Example mappings between ICD-9CM, SNOMED-CT and MedDRA terms

As a result, new mappings can be inferred (e.g., "Anaphylactic shock due to serum, not elsewhere classified" (ICD9) ←matches→ "Anaphylactic transfusion reaction" (MedDRA). However, via reasoning, problematic or incorrect mappings are also found (see Figure 2).

Another important problem is the inclusion of implicit context in the use of a particular term which can result in certain usage contexts with terms that are semantically equivalent in one usage context but not in another [6]. For example, if a generic code for a TNM (TNM Classification of Malignant Tumours) grade is used in local templates of anatomic pathology reports for cancer, knowing the context—breast or prostate cancer for example—the local code for pT shall be bound to the appropriate cancer specific pT code (e.g. either LOINC:44663-3 T classification in Breast tumor or LOINC:44664-1 T classification in Prostate tumor). Taking the context information into account, pT3 value for example has a different value in the context of a breast cancer (Tumor >50 mm in greatest dimension) than in the context of prostate cancer (Extraprostatic extension).

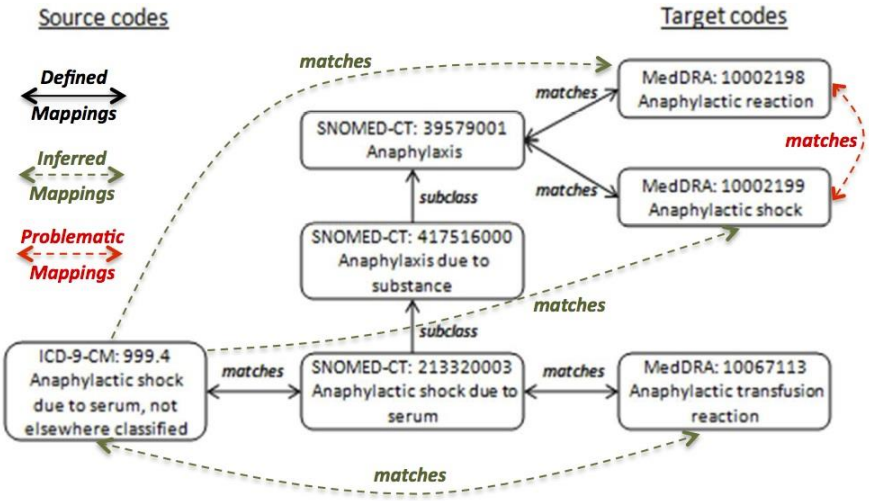

Figure.2. Defined and inferred mappings between ICD-9-CM and MedDRA terms

This paper presents a semantic framework for the representation, evaluation and utilization of terminology mappings. The framework allows terminology mappings to be expressed in a semantic manner, together with the justification for and the provenance information of the mappings. Terminology reasoning, which infers new mappings; and mapping validations, which detects problematic mappings, are also discussed in this paper. In the end, the terminology server of the SALUS project is introduced, where mapping reasoning and validation are applied. A REST API is also established in the terminology server, which allows querying the asserted and inferred mappings. The presentation of mapping relations discussed in this paper are mainly based on SKOS mapping properties [15]. However, the semantic framework presented in this paper is not restricted to using SKOS mapping properties.

2 Semantic Framework Overview

The semantic framework for representing, evaluating and utilizing medical terminology mappings and their justification and provenance is depicted in Figure 3. Its features are:

- Exploitation of available terminology mapping and alignment strategies [1, 7] for finding similarities between source/reference terminologies.
- If the mappings are not expressed semantically, interpret them with semantic expressions.
- Presenting the resulting mappings to terminology experts for further validation.
- Representing the supporting arguments and methods used (i.e. provenance) [8] for defining or finding mappings with a focus on the reason for defining the equivalence relation [9].

- Exploiting reasoners to perform terminology reasoning for finding useful inferred mappings and also to validate both asserted and inferred mappings based on formalized validation schemes and policies [10,11].
- Offering RESTful Web Services, REST APIs and query endpoints for querying and utilizing available terminology mappings—based on their provenance—for different contexts or usages in mind.

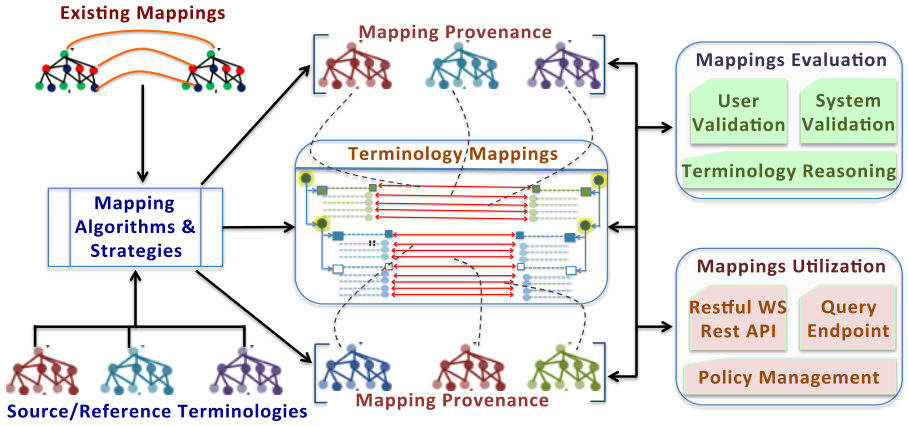

Figure.3. Terminology Mappings Representation, Evaluation and Utilization Framework.

2.1 Terminology Mapping Strategies

There have been various efforts in finding and representing mappings between standard medical terminologies [1,7]. BioPortal, a repository of biomedical terminologies and ontologies, with more than 300 ontologies to date, performs a variety of mapping algorithms and strategies to express different mapping relations with SKOS mapping properties [7]. These include (i) Lexical Mappings using LOOM which are represented via the `skos:closeMatch` property and are generated by performing lexical comparison between preferred labels and alternative labels of terms; (ii) Xref OBO Mappings, represented via the `skos:relatedMatch` property, processed by using Xref and Dbxref properties used by ontology developers to refer to an analogous term in another vocabulary; (iii) Concept Unique Identifier (CUI) Mappings from UMLS, represented via the `skos:closeMatch` property are extracted between term concepts which have a same CUI; (iv) URI-based Mappings, represented via the `skos:exactMatch` property are generated identity mappings between term concepts in different ontologies that are represented by the same URI.

Mappings expressed with SKOS mapping properties have their semantics explicitly expressed, and can be consumed by semantic applications. However, there are also many mappings expressed in non-semantic formats. For example, in the clinical domain, the International Health Terminology Standards Development Organisation, together with the World Health Organization, developed the SNOMED CT to ICD-10 mapping [3], where the mappings are expressed in a spreadsheet. In the OMOP project [4], SNOMED CT to ICD-9-CM mappings are stored in a relational table. Inter-

preting these mappings in a semantic format is therefore required so as to utilize these mappings in semantic web applications. In this context, provenance information is required in order to build the confidence on the interpreted mappings.

2.2 Mapping Provenance and Justification Representations

Existing approaches for exchanging mappings, such as a nanopublications [12] or VoID linksets [13], focus on capturing and modeling the provenance of the mapping, i.e. what has been mapped, by whom, using what tool. What it does not convey is the context in which the assertion was made, i.e. the operational equivalence that is encoded by the mapping. The Open PHACTS project has extended the linkset approach to include such explicit justifications for the mappings [9]. The VoID linkset description is extended with an additional property that conveys why the mapping holds, e.g. the mapped terms have the same preferred label or that they share an identifier. A key advantage of this approach is that existing reasoning mechanisms, e.g. rules for `skos:exactMatch`, can be applied without change since the links themselves are represented using a standard ontology. In addition, existing mappings can be simply extended with additional metadata to enable their reuse; there is no need to regenerate the linking data [9,14].

The semantic framework we propose would require a similar approach to make justifications of terminology mappings explicit. That is, a vocabulary of terms to express the different types of justifications for mappings between terms in terminologies across healthcare and clinical research, such as ICD9, SNOMED CT and MedDRA. For example, taking the above-mentioned term mapping into consideration (see Figure 2), the inferred mapping "Anaphylactic shock due to serum, not elsewhere classified" (ICD9) ←mathches→ "Anaphylactic transfusion reaction" (MedDRA) and its provenance can be represented using the nanopublication schema (see Figure 4).

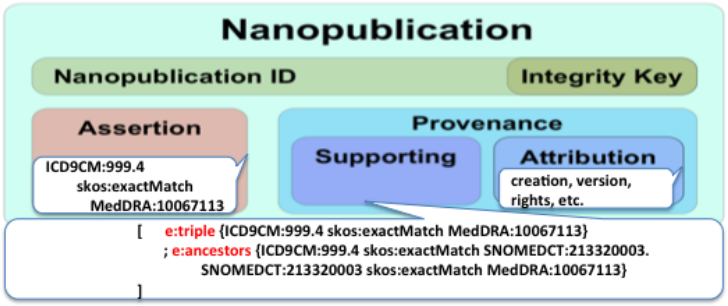

Figure.4. Representing term mapping and its provenance using nanopublication schema

Using the nanopublication schema [12], the inferred mapping are represented as an RDF triple in the `Assertion` graph, and the provenance information related to the mapping assertions are recorded into two categories: (i) `Attribution`, where metadata and context about the mapping can be represented; (ii) `Supporting`, where the justification behind obtaining the recorded mapping assertion are represented. In this

case, the `Supporting` graph includes a 'meta-level' justification trace generated from EYE reasoning engine [18].

2.3 Terminology Reasoning and Mapping Evaluation

Once the existing mappings are formally represented, e.g. using SKOS mapping properties [15], terminology reasoning can be applied to derive new mappings and also to detect both asserted and inferred problematic mappings [11]. Figure 2 shows the inferred mappings between ICD-9-CM and MedDRA by utilizing the existing ICD-9-CM-to-SNOMEDCT and SNOMEDCT-to-MedDRA mappings. However, terminology reasoning could also lead to conclusions that the mapping creators do not intend. Figure 2 shows an example that a match between two MedDRA concepts is inferred because they are both stated as a match to the same SNOMED-CT concept (39579001).

Mapping validations are therefore required to detect erroneous mappings. They are considered as important and often carried out manually by clinicians in the reported terminology mapping projects [1,10]. When the number of mappings is considerably large, quite often a statistically sound sample is extracted, and validations are applied on the sample only. Therefore, the main goal of such validations is to monitor the quality of the mappings, rather than correcting false mappings. In addition, as pointed out in [1], existing terminology mappings need to be updated when related terminologies evolve. It is therefore important to have an automated mapping validation workflow that can detect problematic mappings without human involvement.

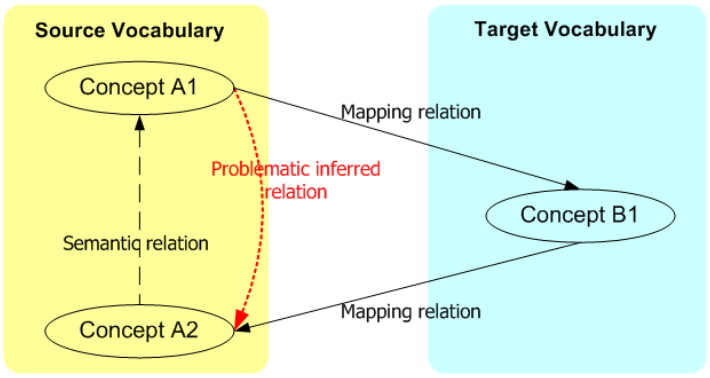

Figure. 5. Basic Pattern of Problematic Mappings Detection

Figure 5 presents the basic pattern of problematic mappings which may produce unintended mappings [11]. When the mapping relations can be considered as transitive, it is possible to infer a relation (the red dotted line in Fig. 5) between two concepts in a same terminology system. If such an inferred relation is not stated in (or cannot be inferred from) that terminology system, then inferring such a relation via mapping relations is considered as vocabulary hijacking, e.g. the problematic mapping displayed in Figure 2. In addition, if such an inferred relation is contradictory to

any existing relation, e.g. the semantic relations displayed as black dashed line in Figure 5, it would be considered as a conflict, and the related mappings need to be corrected.

When the mapping relations and semantic relations in Figure 5 are represented in a formal way, e.g. using SKOS properties, it is possible to detect the problematic patterns in an automated way. The SKOS mapping rules [17] define the required validation rules to detect these problematic patterns. The SKOS validation rules have been implemented and are used on the SALUS terminology server to detect violations. They have proven to be effective.

2.4 Terminology Mapping Querying and Utilization

Once terminology mappings are represented together with provenance information, utilizing them in interoperating applications is another challenge. The SALUS Terminology Server addresses this challenge and provides a RESTful interface for querying term mappings [16]. The SALUS Terminology Server is built upon a triple-store and inherently supports RDF. Term representations are utilized mainly by using SKOS [15]. REST methods return query results according to this semantic modeling of the terms.

The SALUS terminology server stores the concepts of a list of terminology systems, e.g. ICD9CM, ICD10GM, SNOMEDCT, etc., as well as a set of mappings among different terminologies. REST methods provided by SALUS Terminology System can be divided into three categories.

1. Methods that directly return data according to the RESTful principles.
2. Text based search methods that use the labels of the terms for text matching.
3. Query methods for available mappings. The SALUS inference engine outputs `skos:exactMatch` mappings and new SALUS specific mapping predicates such as `exact-or-narrow` and `exact-or-broad`. The REST API can be easily extended with new predicates and inference results.

Terminology systems stored in the SALUS terminology server can be identified either by their unique Object Identifiers (ISO OIDs) or by their labels. A terminology system may also have more than one label. For example, the MedDRA (Medical Dictionary for Regulatory Activities Terminology) terminology (Version 13.0) in the terminology server can be identified by its OID (2.16.840.1.113883.6.163), or by its label (MDR or MedDRA).

Queries to the terminology mappings need to specify the mapping relation, the identifier of source and target terminology system (either by OID or label), and the code in the source terminology system. Below are the query templates:

/terminologyserver/mappings/{mapping-relation}/oid/{oid}/{code}?toid={toid}

/terminologyserver/mappings/{mapping-relation}/label/{label}/{code}?tlabel={tlabel}

3 Results

Table 1. Terminology reasoning and mapping validation in the SALUS terminology server

Mapping	Source	Target	Relations in Vocabularies*	Mapping relations		Number of Mappings	Problematic Patterns
1	SNOMED-CT	ICD-10	1,839,401	Imported from: Crossmap	skos:exactMatch exactOrNarrow	16,710 11,346	104,761 -
				Inferred:	exactOrNarrow**	16,838	-
2	MedDRA	SNOMED-CT	1,886,575	Imported from: OntoADR	skos:exactMatch	10,648	1,790
				Inferred:	exactOrNarrow	453,615	-
3	MedDRA	ICD-10	121,032	Inferred from: Mapping 1, 2	skos:exactMatch exactOrNarrow	3,072 136,144	3,331 -
4	SNOMED-CT	ICD-9CM	1,845,688	Imported from: OMOP	exactOrNarrow	16,819	-
				Inferred:	exactOrNarrow	31,862	-
5	MedDRA	ICD-9CM	174,493	Inferred from: Mapping 2, 4	exactOrNarrow	270,094	-

* Relations in Vocabularies refers to the skos:broaderTransitive relation stated inside the source or target vocabularies.
** exactOrNarrow is a mapping relation used in SALUS to denote a mapping which is either exact or narrower.

Table 1 shows the mappings used in the SALUS terminology server [16], as well as the output of terminology reasoning and mapping validation. Mappings 1, 2 and 4 are created based on mappings developed by existing projects. In particular:

- Mapping 1: US NLM provides mapping between SNOMED CT to ICD-10 to support semi-automated generation of ICD-10-CM codes from clinical data encoded in SNOMED CT for reimbursement and statistical purposes. This is a result of CrossMap Project by IHTDSO and WHO
- Mapping 2: IMI PROTECT project created an ontology called OntoADR, which also presented the correspondence between MedDRA and SNOMED CT codes
- Mapping 4: OMOP project provides mappings of a selected subset of ICD9CM and ICD10CM codes to SNOMEDCT Clinical Findings.

These existing mappings are interpreted into mappings represented either as `skos:exactMatch`, or the SALUS `exactOrNarrow` relation. Based on the imported mapping relations, as well as semantic relations (`skos:broaderTransitive`) stated inside of the related terminologies, some additional mapping relations (mainly the `exactOrNarrow` relation) are inferred as well. In addition, by using SNOMED-CT as a bridge terminology, the MedDRA-to-ICD10 mappings (mapping 3) are derived from mappings 1 and 2, and the MedDRA-to-ICD9CM mappings (mapping 5) are derived from mappings 2 and 4.

The SKOS mapping validation rules described in [16,17] are applied on the mappings listed in Table 1 to detect problematic mappings. The column 'Detected Patterns' in Table 1 shows the number of detected problematic patterns. It can be observed that there are many problematic patterns detected. This is because wrong assumptions were made in interpreting the original mappings stored in spreadsheets or

in databases when converting them to RDF by assigning either `skos:exactMatch` or `exactOrNarrow` property.

The validation study finds it is difficult to interpret the existing mappings into correct SKOS mappings by a third. In case such an interpretation is requested, it is important to provide provenance information together with the mappings so as to make the interpretation process explicit. It is also worth mentioning that when mappings are expressed with mapping properties with a strong semantic commitment (e.g. `skos:exactMatch`, `skos:broadMatch`), more entailments can be achieved, however, it is also more likely to violate the SKOS mapping constraints. While the mappings that are expressed with only a weak semantic commitment, e.g. the mappings in Bioportal are mostly expressed with `skos:relatedMatch` and `skos:closeMatch`, they are less likely to violate the SKOS mapping constraints. However, their usage is also much more limited due to the weak semantic commitment.

In order to make the existing mappings reusable over the semantic web, it is extremely important that the communities who created the mappings also provide their mappings in RDF using standard ontologies to represent their mappings (e.g. SKOS). By using RDF properties, the mapping relations are more explicitly stated compared with text description. On top of this, we realized that the mapping relationships provided by external sources are not always simple 1-1 relations between the source and target vocabulary; 1-n mappings and n-m mappings also exist. Meanwhile, some mappings require additional conditions to conduct a mapping. In the original mappings of the Crossmap project, around 25% of their mappings are associated with conditions. For example, the SNOMED-CT code “430556008” for “Malignant neoplasm of genital structure” is mapped to ICD-10 code “C57.9” if the patient is female, and to “C63.9” if the patient is male. Those mappings are difficult to be expressed by SKOS mappings, and would better be treated as mapping rules. For applications that intend to infer additional mappings based on existing ones, it is important to notify that the SKOS mapping properties are not transitive (except for the `skos:exactMatch`).

The justification and provenance information is not yet stored in the SALUS terminology server. However, it is possible to provide provenance information for each asserted mapping set, together with the reason for the mapping, as either a VoID description or a nanopublication. In addition, provenance information for each inferred mapping could also be asserted in the SALUS terminology server—following the example guideline suggested in Figure 4.

4 Conclusion

The role of medical terminologies is vital to achieve computable semantic interoperability (CSI) in clinical informatics. In this paper, we show the challenging nature of mapping utilization among different terminologies. The introduced semantic framework has been built upon existing terminology mappings to (i) infer new mappings for different CSI use cases, (ii) present provenance of the mappings together with the

justification information—an important problem for term mapping utilization, and (iii) perform mapping validation in order to show that inferred mappings can be erroneous.

The semantic framework enables a more collaborative semantic landscape with providers and consumers of terminology mappings. It can also be the basis for additional services, such as usage data and feedback mechanisms for the providers of mappings and of the source and targets terminologies. Making the use, and reuse, of terminology mappings visible could enable new funding and business models in a sustainable semantic landscape across clinical care and clinical research domains.

The terminology mappings and validations in this paper are expressed with SKOS, as it is widely used in expressing controlled vocabularies. The semantic framework itself is not restricted to using SKOS, but is open to other ontologies that are capable of expressing mapping relations.

References

1. Saitwal H, Qing D, et al. Cross-terminology mapping challenges: A demonstration using medication terminological systems. *Journal of Biomedical Informatics*, **45(6)**:1217(2012).
2. Erturkmen, G.B.L., et al. SALUS: Scalable, Standard based Interoperability Framework for Sustainable Proactive Post Market Safety Studies. MIE 2012, Pisa-Italy, 26-29 August 2012.
3. Williams, A.J. et al. Open PHACTS: semantic interoperability for drug discovery. *Drug discovery today*, 17(21-22), pp.1188–98 (2012).
4. Ouagne D., et al. The Electronic Healthcare Record for Clinical Research (EHR4CR) information model and terminology. Studies in health tech. and informatics (2012), 534-8.
5. A Study of Terminology Mapping in SALUS Project. http://www.srdc.com.tr/projects/salus/blog/?p=241, Nov 15, 2013.
6. Rector A., Qamar R., Marley T. Binding ontologies and coding systems to electronic health records and messages. *Applied Ontology*, **4(1)**, (2009), 51-69.
7. Salvadores, M., Alexander, P. R., et al. BioPortal as a Dataset of Linked Biomedical Ontologies and Terminologies in RDF. Semantic Web Journal, **4(3)** (2013), 277-284.
8. McGuinness D. L., Silva P. P. Explaining Answers from the Semantic Web: The Inference Web Approach. International Semantic Web Conference, (2003).
9. C. Batchelor, et al. Scientific lenses to support multiple views over linked chemistry data. In International Semantic Web Conference (2014). To appear.
10. White S, Paoin W et al. SNOMED CT to ICD-10 Map Phase 1 Content Validation Report. http://www.ihtsdo.org/fileadmin/user_upload/Docs_01/SNOMED_CT/Development/Content_Validation_Report_20130327.pdf, Last visited on January 2014.
11. Sun H, De Roo J, Twagirumukiza M, et al. Validation Rules for Assessing and Improving SKOS Mapping Quality. arXiv preprint arXiv:1310.4156, (2013).
12. Nanopublication Guidelines. http://www.nanopub.org/guidelines/, 15 December 2013.
13. Describing Linked Datasets with the VoID Vocabulary. http://www.w3.org/TR/void/, W3C Note, 3 March 2011.
14. Brenninkmeijer, C.Y.A. et al. Computing Identity Co-Reference Across Drug Discovery Datasets. In Semantic Web Applications and Tools for Life Sciences (2013).
15. SKOS Simple Knowledge Organization System Reference. http://www.w3.org/TR/skos-reference/, W3C Recommendation, 18 August 2009

16. Depraetere, K., Erturkmen, G.B.L. Deliverable 4.4.2-SALUS Semantic Mediation Framework–R2. http://www.srdc.com.tr/projects/salus/public-deliverables, April 30, 2013.
17. SKOS Mapping Validation Rules. http://eulersharp.sourceforge.net/2003/03swap/skos-mapping-validation-rules. Last visited on July 2014.
18. Jos De Roo. Euler Yap Engine, http://eulersharp.sourceforge.net/2003/03swap/eye-note.txt Last visited on July 2014.