Stud Health Technol Inform (Proc Medinfo) 2013:(in press).

Leveraging Terminological Resources for Mapping between Rare Disease Information Sources

Bastien Rance[a], Michelle Snyder[b], Janine Lewis[b], Olivier Bodenreider[a]

[a] *U.S. National Library of Medicine, National Institutes of Health, Bethesda, Maryland, USA*
[b] *Genetic and Rare Diseases Information Center, ICF International, Rockville, Maryland*

Abstract

Background*: Rare disease information sources are incompletely and inconsistently cross-referenced to one another, making it difficult for information seekers to navigate across them. The development of such cross-references established manually by experts is generally labor intensive and costly.* ***Objectives****: To develop an automatic mapping between two of the major rare diseases information sources, GARD and Orphanet, by leveraging terminological resources, especially the UMLS.* ***Methods****: We map the rare disease terms from Orphanet and ORDR to the UMLS. We use the UMLS as a pivot to bridge between the rare disease terminologies. We compare our results to a mapping obtained through manually established cross-references to OMIM.* ***Results****: Our mapping has a precision of 94%, a recall of 63% and an F_1-score of 76%. Our automatic mapping should help facilitate the development of more complete and consistent cross-references between GARD and Orphanet, and is applicable to other rare disease information sources as well.*

Keywords: rare diseases; terminologies; interoperability

Introduction

One common issue experienced by rare diseases patients, their families and health professionals is the lack of information about a specific disorder [1]. Several comprehensive sources of information about rare diseases have emerged in the past decade in the U.S. and in Europe, including the Genetic and Rare Diseases Information Center (GARD) [http://www.rarediseases.info.nih.gov/GARD/], the Rare Disease Database created by the National Organization for Rare Disorders (NORD) [http://www.rarediseases.org/] and Orphanet [http://www.orpha.net/]. The Online Mendelian Inheritance in Man (OMIM) is the oldest of these resources and provides extremely detailed information about human genes and genetic phenotypes, but is intended primarily for use by health professionals.

These information sources are partially cross-referenced to standard medical terminologies and among themselves. One example of relatively well cross-referenced disease is *neurofibromatosis type 2*. In Orphanet, it is cross-referenced to OMIM and to several standard medical terminologies, including ICD-10, MeSH, MedDRA, SNOMED CT and Unified Medical Language System (UMLS). GARD provides cross-references to OMIM, NORD and Orphanet. In contrast, no cross-reference to OMIM is provided by either source for the disease *hemifacial microsomia*, for which GARD nonetheless provides a mapping to Orphanet. As illustrated in these examples, such cross-references are not necessarily consistent and are often incomplete, making it difficult for users to navigate across these resources. Moreover, the development of such cross-references established manually by experts is generally labor intensive and costly, and therefore difficult to maintain over time when resources are updated.

The objective of this investigation is to develop an automatic mapping between two rare disease information sources, GARD and Orphanet, by leveraging terminological resources, especially the UMLS. Such an automatic mapping is expected to facilitate the development and maintenance of complete and consistent cross-references between the two information sources.

Background

Ontology Alignment

This investigation is in the general framework of ontology alignment (also called ontology mapping or matching). Exhaustive reviews of existing work can be found in [2] and [3].
In addition to the usual techniques (e.g., lexical mapping, semantic filtering), like [4], we also use mappings to a reference ontology in order to infer the mapping between our source and target ontologies.

Aligning Rare Disease terminology to the UMLS

[5] investigated the need for automatic methods to assist expert in the process of mapping rare disease terms to the UMLS, with application to the Orphanet terminology. Their method relies mostly on an aggressive normalization and is reported to have a high precision (94.6%), based on the manual evaluation of 2476 equivalent mappings. This method also supports the creation of partial mappings, of which they do not report the performance.

The Unified Medical Language System (UMLS)

The Unified Medical Language System® (UMLS®) is a terminology integration system developed at the National Library of Medicine. The UMLS Metathesaurus® integrates more than 160 biomedical vocabularies. Synonymous terms from the various source vocabularies are grouped into one concept. Additionally, the Metathesaurus records the relations asserted among terms in the source vocabularies, including hierarchical, associative and mapping relations. Version 2012AA of the UMLS is used in this study. This version contains approximately 2.6 M concepts and 40 M relations.
The integration process in the UMLS uses a semi-automatic method based on the normalization of terms. Terms with the same normalization are candidates to being grouped into a single UMLS concept and are then manually reviewed by UMLS editors. The UMLS normalization process is illustrated

in ***Table 1***, using the term *Fried's tooth and nails syndrome* as an example.
Each UMLS concept is categorized with at least one semantic type from the UMLS Semantic Network. Groupings of semantic types provide an easy way of selecting all concepts from a given subdomain of medicine, e.g., all disorders with the semantic group *Disorders*.

Table 1 - Normalization process in the UMLS

Step	Results
Original string	Fried's tooth and nails syndrome
Remove genitive	Fried tooth and nails syndrome
Remove stop words	Fried tooth nails syndrome
Lowercase	fried tooth nails syndrome
Strip punctuation	fried tooth nails syndrome
Uninflect	fried tooth nail syndrome
Sort words	fried nail syndrome tooth

Materials and Methods

Materials

Rare Diseases Terms from the Office of Rare Diseases Research (ORDR)

The Genetic and Rare Diseases Information Center (GARD) is a collaborative effort of two agencies of the U.S. National Institutes of Health to help people find useful information about genetic conditions and rare diseases. One of these agencies, the Office of Rare Diseases Research (ORDR) of the National Center for Advancing Translational Sciences (NCATS), publishes a list of rare diseases, the "Rare Diseases and Related Terms". This list comprises 6,316 rare disease concepts (6,316 preferred terms and 12,627 synonyms). The rare disease concepts correspond to diseases for which information requests have been made or diseases that have been suggested as being rare. GARD provides extensive information about 1100 of these diseases. The purpose of the Rare Diseases and Related Terms list is to facilitate the distribution of information. In addition to this list of disease terms, GARD has shared with us the cross-references they have established to OMIM.

Rare Diseases Terms from Orphanet

Orphanet is "the reference portal for information on rare diseases and orphan drugs, for all audiences". Orphanet is based in Europe and provides an inventory of rare diseases and drugs, as well information about rare diseases with the goal of helping to improve the diagnosis, care and treatment of patients with rare diseases. In practice, Orphanet provides information about 6,578 rare diseases. Orphanet diseases are organized into a Directed Acyclic Graph. In the Orphanet database, diseases are linked to external reference terminologies, such as ICD10 and OMIM. The Orphanet list of rare diseases comprises 6,578 concepts (6,578 preferred terms and 7,552 synonyms). Additionally, Orphanet has established cross-references between rare disease concepts and OMIM, and various reference terminologies including ICD10, MeSH, SNOMED CT, MedDRA and the UMLS.

Methods

Our method to find mappings between rare diseases terminologies can be summarized as follows. First we mapped the rare disease terms from Orphanet and the ORDR list to the UMLS. We used the UMLS as a pivot to bridge between the rare diseases. Finally we evaluated the quality of the results.

Mapping rare disease terms to the UMLS

In order to map rare diseases terms from the two information sources to UMLS concepts, we use a series of increasingly aggressive methods depicted in *Figure 1*. First, we attempt to find an equivalent concept through string match. If it fails, we attempt to find a broader concept using word subsets. For each strategy, the strictness of the matching criteria can be relaxed from exact match, to UMLS normalization, to extended normalization. Regardless of the mapping strategy, we apply semantic constraints to all mappings in order to keep mappings to the realm of diseases. We also ignore terms corresponding to acronyms, because of their inherent ambiguity.

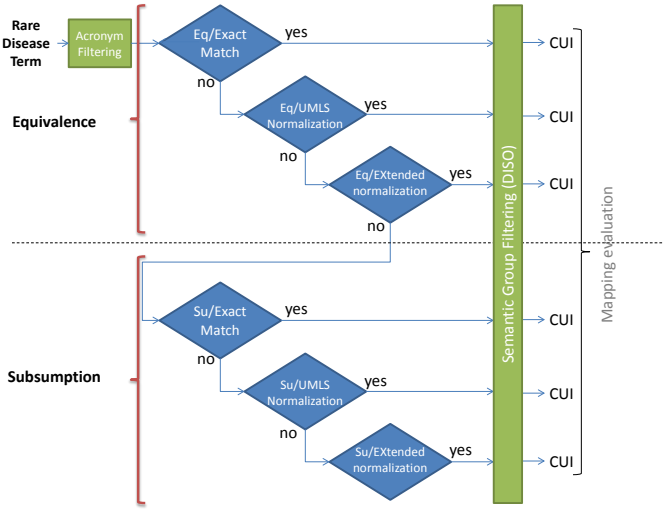

Figure 1 - Mapping Rare Disease terms to the UMLS

a) Finding equivalent concepts through string match.
The least aggressive mapping strategy is to find equivalent concepts through string match (Eq).
Exact match (EM). We first try to match the rare disease term to a synonym in the UMLS using an exact match strategy against the UMLS. For example, the ORDR term *Verloove Vanhorick Brubakk syndrome* maps to the UMLS concept C1859082 through exact match.
UMLS normalized match (UN). Then, we normalize all rare disease terms using the normalization function provided by the UMLS and attempt to match them against similarly normalized terms in the UMLS. For example, the Orphanet term *Infantile symmetrical thalamic degeneration* maps to the UMLS concept *Symmetrical infantile thalamic degeneration* [C2931220] after UMLS normalization.
Extended normalization (EX). In some cases, the UMLS normalization is too conservative and fails to identify an existing concept of the UMLS. We extended the UMLS normalization based on [5]. More specifically we used three additional steps: (i) transforming Roman numerals into Arabic numerals (e.g., *iii* becomes *3* and *ixc* becomes *9c*), (ii) extending the stop word list with domain specific, inconsistently used words such as 'type' and 'syndrome'; and (iii) normalizing the karyotype formats (e.g, *48, XXXY* becomes *XXXY*). We apply our extended normalization to all rare disease terms before attempting a match against normalized terms in the UMLS. (We assume that rare disease names in the information sources under investigation exhibit more variability than those in reference terminologies, which are already normalized in the UMLS.) For example, the Orphanet term *Familial restrictive cardiomyopathy type 2* maps to the OMIM concept *CARDIOMYOPATHY, FAMILIAL RESTRICTIVE, 2* in UMLS after extended normalization (during which type is removed).

b) Finding broader/narrower concepts through word subsets.

If no results can be found through string match, we attempt to find a broader or narrower concept in the UMLS, i.e., the source and target concepts are in a subsumption relationship (Su). To this end, we leverage lexical semantics principles and assume that the set of words in the name for the broader concept will a proper subset of the set of words in the name for the narrower concept. Like string matching, mapping through word subsets can be more or less strict, depending on whether the word subsets are derived from the original terms or from normalized terms. For example, the Orphanet term *Ehlers-Danlos syndrome, classic type* maps (through exact match) to two narrower concepts in the UMLS, whose terms contain all the words of the original rare disease term plus some other words: *Ehlers-Danlos Syndrome, Severe Classic Type* (a synonym for *Ehlers-Danlos syndrome type 1* [C0268335]) and *Ehlers-Danlos syndrome, mild classic type* (a synonym for *Ehlers-Danlos syndrome type 2* [C0268336]).

Filtering of acronyms. Due to the ambiguity of acronyms, we ignore mappings obtained solely by matching to an acronym. For example, the ORDR term *BBS* is excluded from our processing. In addition to *Bardet-Biedl syndrome*, it would also be mapped (wrongly) to *Berlin Breakage Syndrome*.

Semantic constraints. Because terms from ORDR and Orphanet are all expected to be names for (rare) disorders, we restrict the UMLS concepts mapped to to disorder concepts. In practice we only consider mappings to concepts from the Semantic Group *Disorders* (including such semantic types as *Disease or Syndrome* and *Congenital Abnormality*). This simple filter provides some level of word sense disambiguation. For example, the source term *NF2* can be mapped to both a disease (*neurofibromatosis type 2*) or to a gene (*NF2*, on chromosome 22, whose mutation causes *neurofibromatosis type 2*). Constraining the mapping to disorder concepts helps us avoid a wrong mapping to the gene concept. There might be residual ambiguity, however, when a source term maps to several disorder concepts.

Mapping terminologies using the UMLS as a pivot terminology

After all terms from ORDR and Orphanet have been mapped to the UMLS, it is possible to use the UMLS as a pivot terminology to derive a mapping between ORDR and Orphanet through the UMLS. When the ORDR and Orphanet terms map to the same UMLS concept, we can derive a direct mapping between the two sources. In contrast, when the ORDR and Orphanet terms map to different, but hierarchically related UMLS concept, we can derive an indirect mapping between the two sources. These two mapping situations are illustrated in ***Figure 2***. By including indirect mappings (through subsumption relations), we know we take the risk of generating false positive mappings. However, we want to investigate the effect of this more aggressive strategy on the performance of the mapping algorithm.

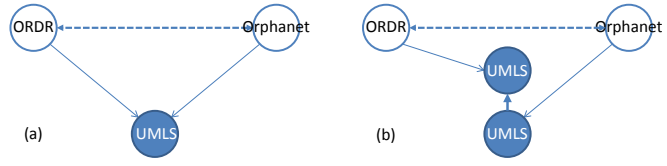

Figure 2 - Direct (a) and indirect (b) mappings between ORDR and Orphanet terms through UMLS concepts

Examples of direct mappings include the mapping between the ORDR concept *Propionic acidemia* and the Orphanet concept *Propionicacidemia* through the UMLS concept C0268579, for which both terms are names. The ORDR concept *Paris-Trousseau thrombocytopenia* maps to UMLS concept C1956093, while the Orphanet concept *Paris-Trousseau syndrome* maps to C0795841. However, since this two concepts are hierarchically related in the UMLS – C0795841being broader than C1956093 – an indirect mapping is established between the two rare disease concepts. More precisely, *Paris-Trousseau syndrome* is broader than *Paris-Trousseau thrombocytopenia*.

Implementation. We leveraged the UMLS Terminology Services (UTS) API 2.0 to identify UMLS concepts corresponding to rare disease terms and to acquire UMLS information about concepts. Information about mapping to UMLS and to OMIM was loaded into a triple store. We used rules to automatically derive the mappings between rare disease concepts through OMIM and through UMLS.

Evaluating the quality of the mapping through UMLS

Both ORDR and Orphanet provide cross-references to OMIM established by experts for a majority of their rare disease terms. We take advantage of these cross-references to the same external source for the evaluation of our automatic mapping through the UMLS. Since we use the mapping to the same OMIM concept as evidence of a mapping between two rare disease concepts, we restrict the reference to the set of rare disease concepts that are mapped to at least one OMIM term. From the perspective of our mapping algorithm, we consider that a mapping was found through the UMLS if any of the mapping strategies succeeded.

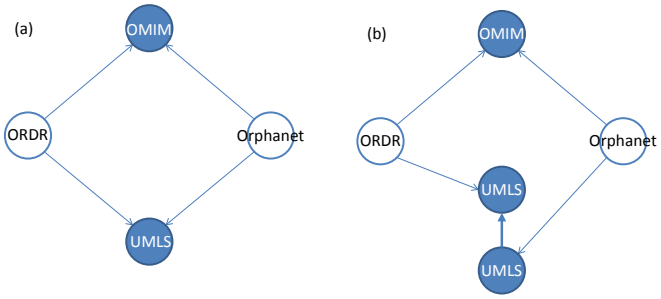

Figure 3 - Comparison of the automatic mapping to the mapping through OMIM

As illustrated in ***Figure 3***, our assumption is that when two concepts from ORDR and Orphanet are associated with the same OMIM term, we should also find a mapping between them, direct or not, through the UMLS. Cases (a) and (b) are considered true positives. True negative cases are not known because, as mentioned earlier, this evaluation is restricted to those concepts from ORDR and Orphanet that are cross-referenced to OMIM.

We also investigated discrepancies between the mappings identified through OMIM (reference) and our automatic mappings through UMLS (***Figure 4***). We consider false positives the cases where there is no mapping through OMIM, but a mapping through the UMLS. Conversely, the cases where there is a mapping through OMIM, but a mapping through the UMLS are false negatives.

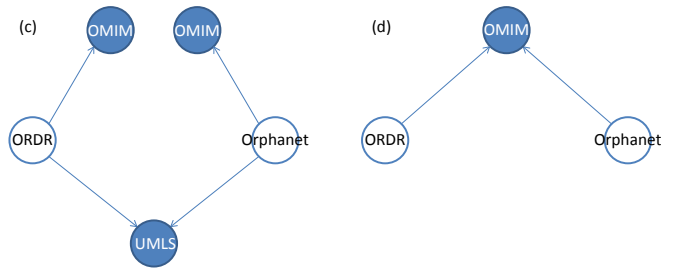

Figure 4 – Discrepancies between the mappings identified through OMIM and UMLS: false positives (c) and false negatives (d)

We evaluate the performance of our mapping algorithm against the reference mapping to OMIM using the classic precision, recall and F_1-score for each of the strategies. The F_1-score is the weighted harmonic mean of precision and recall.

Manual review of the false positives

Because the mapping of the sources to OMIM do not claim to be complete, one of the authors (BR) manually reviewed the false positive mappings discovered by the automatic method in order to assess if they could be explained by errors in cross-references in the sources.

Results

Mapping of ORDR and Orphanet concepts to UMLS

Of the 6316 ORDR concepts, 5361 (85%) could be mapped to a UMLS concept through at least one of their terms. Similarly, 4451 of the 6578 Orphanet concepts (68%) could be mapped to a UMLS concept.

As shown in *Table **2***, a majority of the mappings are equivalence mappings, and simple techniques, such as exact match and normalization, contribute most of the mappings.

Table 2 - Contribution of each technique in the mapping to the UMLS

		ORDR	Orphanet
Eq	EN	4744	3163
	UN	397	826
	XN	22	99
Su		1153	363

Mapping between ORDR and Orphanet through UMLS

Overall, we derived an automatic (direct) mapping in 4235 cases between the 5361 ORDR concepts (79%) and 4451 Orphanet concepts (95%).

Performance evaluation

The 3396 ORDR concepts and 3782 Orphanet concepts having a mapping to OMIM constitute the reference for our evaluation, since the mapping through OMIM is used as the reference.

As mentioned earlier, our main focus is on the direct mappings. We also report the performance for all (direct and indirect) mappings. Results are summarized in ***Table 3***.

Our direct mapping through OMIM was able to identify 2155 of the 3479 pairs of ORDR and Orphanet concepts associated through OMIM, and identified 241 additional associations.

In terms of the standard metrics, the performance of our (direct) mapping algorithm is as follows: recall: 61.94%, precision: 89.94%, F_1: 73.36%. As expected, extending the mapping algorithm to the more aggressive technique (indirect mapping) increases recall at the cost of severely decreasing precision (recall: 68.41%, precision: 49.43%, F_1: 57.39%), which may not be useful in practice.

Table 3 – Evaluation of the the automatic mapping through UMLS against the reference mapping to OMIM, and specific contribution of the direct mappings.

Direct only

		Mapping through OMIM	
		Yes	No
Mapping through the UMLS	Yes	2155	241
	No	1324	

All

		Mapping through OMIM	
		Yes	No
Mapping through the UMLS	Yes	2380	2435
	No	1099	

Manual Review

We manually reviewed the 207 direct mappings obtained through the UMLS but not corroborated by a mapping through OMIM ("false positives"). In 50 cases, we classified the mappings as correct (suboptimal mapping to OMIM in the reference). In 54 cases, the OMIM concepts cross-linked to were close and our mapping through UMLS is acceptable. Finally, 103 mappings through UMLS were incorrect. Getting credit for these 104 cases (excluding only the 103 wrong mappings) would slightly increase the performance of our mapping algorithm (recall: 63.05%, precision: 94.24%, F_1: 75.55%).

Discussion

In this section, we discuss the practical significance of our findings, the technical significance of our approach and some of its limitations.

Findings and practical significance

Findings. We showed that it is possible to create an automatic mapping between ORDR and Orphanet. This mapping covers 80-95% of the concepts in each source and its performance is reasonably good, although recall is relatively low.

Prospective use. In practice, the automatic mapping can be implemented easily and updated frequently. This mapping could be used to support the original development and maintenance of a cross-reference between GARD and Orphanet. In particular, compared to the present situation, this automatic mapping could assist domain experts in producing a complete and consistent cross-reference between GARD and Orphanet, which would help information seekers navigate across these two information sources more effectively.

Harmonization of rare disease terminological resources. In addition to establishing a cross-reference between two information sources, our mapping would also help harmonize their terminological resources. In fact, each source uses some synonyms that are not found in the other source. On the basis of equivalences found through these mappings, we estimated that 3024 Orphanet synonyms could be added to ORDR terms, and 6219 ORDR synonyms could be added to Orphanet terms. The average number of synonyms per concept would increase from 3.0 to 3.47 in ORDR and from 2.15 to 3.09 in Orphanet.

Improving cross-references to OMIM. As we mentioned, cross-references to OMIM are incomplete in GARD and Orphanet. In cases where a mapping is found through UMLS, but only one of the sources is cross-referenced to OMIM, a cross-reference to OMIM can be inferred for the other source in some cases. For equivalence mappings, the other source

should be cross-referenced to the same OMIM concept. In other words, the equivalence mapping obtained through UMLS helps identify missing cross-references with high confidence. Considering only the relations identified through exact or normalized match, our method identifies 297 missing cross-references to OMIM in GARD, and 212 in Orphanet.

Two authors (MS and JL) have reviewed the OMIM suggestions associated with 48 ORDR concepts (165 OMIM cross-reference predictions). 36 concepts (77 predictions) had a least one mapping considered equivalent (18) or related (18). Most errors are due to the incorrect mapping of one single ORDR term to a UMLS concept, resulting in 50 incorrect predictions.

Mapping of non-genetic diseases. Unlike OMIM, the UMLS is not restricted to genetic diseases. Since rare diseases are not necessarily of genetic origin, the mapping through UMLS yields additional results compared to the mapping through OMIM. We showed that 4235 pairs of ORDR and Orphanet concepts are associated through UMLS, while only 3479 are associated through OMIM.

Generalization. A similar approach could be used to create cross-references with other rare disease information sources, including OMIM, NORD and the Genetic Home Reference [http://ghr.nlm.nih.gov/]. Applications beyond rare diseases are possible, but may require customization of the extended normalization to the specific lexical forms used in a given subdomain.

Technical significance

Extended normalization. Domain-specific normalization has already been suggested for specific types of biomedical entities, whose names exhibit specific variation (e.g., for drugs [6]). Arguably, some aspects of the domain-specific normalization we propose here (e.g. replacing Roman numerals with Arabic numerals, karyotype normalization) are not specific to rare disease names and could be extended to the broader domain of disorders. Extended normalization could be integrated into the UMLS lexical programs. In practice, it provided modest benefits in this study, and would have to be carefully evaluated before a broader application to the UMLS is performed.

Using hierarchical relations. We used hierarchical relations from the UMLS to reconcile differences in granularity between concepts from the two rare disease terminologies. While this indirect mapping approach increased the recall of our automatic method by 3%, it also generated an important set of potential mappings of interest. In this study, we considered all possible hierarchical relations, using a transitive closure. This approach could be refined to allow only close hierarchically related concepts to contribute to the mapping.

Confidence levels. Our approach to mapping between rare disease information sources uses a sequence of increasingly aggressive techniques. We first attempt to find a mapping directly through the UMLS, before attempting to relate UMLS concepts mapped to through hierarchical relations. Moreover, in the mapping of rare disease terms to the UMLS, we also use increasingly aggressive techniques, first attempting to find equivalent concepts (with various levels of normalization), before we resort to controlled approximate matches through word subsets. Each step in the mapping process can be associated with a level of confidence. In general, the confidence level in a mapping between ORDR and Orphanet concepts is a function of the confidence level of the mapping of each concept to UMLS, as well as the confidence level in the mapping through UMLS (i.e., direct vs. indirect). For example, a mapping between ORDR and Orphanet through the exact match of the two terms to the same UMLS concept will have the highest level of confidence, whereas the introduction of normalization or the use of approximate matching techniques on one side or both will lower the confidence we may have in the mapping.

Limitations

The automatic mapping approach between GARD and Orphanet concepts presented here still requires validation by rare disease domain experts before it can be published. Although its precision is acceptable, it still generates a number of false positives and would be best used to facilitate the work of experts.

The evaluation of the performance of our algorithm relies on the cross-reference to OMIM provided by each source. However, the exact nature of the cross-reference (equivalence or broader/narrower) is not specified. Therefore, although two rare disease concepts are cross-referenced to the same OMIM concept, they are not necessarily equivalent unless the cross-reference to OMIM on each side denotes equivalence.

Conclusion

In this study we presented an automatic approach to mapping between rare disease information sources. We relied on the UMLS as a pivot terminology and used an extended normalization technique to improve the coverage of the method. Compared to a mapping derived from manually curated reference to OMIM, our precision is 90% and recall 62%. This automatic mapping can facilitate the development of cross-references between, and ultimately the interoperability of, GARD and Orphanet. Additional benefits include enriching and harmonizing the underlying terminological resources.

Acknowledgments

This research was supported in part by the Intramural Research Program of the National Institutes of Health, National Library of Medicine (NLM), National Center for Advancing Translational Sciences (NCATS), and National Human Genome Research Institute (NHGRI).

References

[1] Budych K, Helms TM, Schultz C. How do patients with rare diseases experience the medical encounter? Exploring role behavior and its impact on patient-physician interaction. Health Policy. 105(2-3):154-64, 2012.

[2] Euzenat J, Shvaiko P. Ontology Matching. Berlin Heidelberg (DE): Springer-Verlag; 2007.

[3] Bellahsene ZB, Angela; Rahm, Erhard. Schema Matching and Mapping: Springer-Verlag; 2011.

[4] Groß, A, Hartung, M, Kirsten, T, Rahm, E. Mapping Composition for Matching Large Life Science Ontologies. 2nd International Conference on Biomedical Ontology (ICBO 2011) 2011.

[5] Brandt MM, Rath A, Devereau A, Aymé S. Mapping Orphanet Terminology to UMLS. In: Peleg M, Lavrač N, Combi C, eds. Springer-Verlag, Heidelberg, AIME 2011, LNAI 6747, 2011; pp194-203.

[6] Peters L, Kapusnik-Uner J, Bodenreider O. Methods for Managing Variation in Clinical Drug Names. In: Proc of AMIA 2010; pp637-41, 2010

NLM, NIH, Blg 38A, 8600 Rockville Pike, 20894 Bethesda MD, USA

Corresponding author: olivier@nlm.nih.gov