Ontology Matching for the Laboratory Analytics Domain*

Ian Harrow[1], Thomas Liener[1], and Ernesto Jiménez-Ruiz[2,3]

[1] Ontologies Mapping Project, Pistoia Alliance, USA
[2] City, University of London, United Kingdom
[3] SIRIUS, Department of Informatics, University of Oslo, Norway

1 Introduction

The Pistoia Alliance was established ten years ago to promote innovation by industry through pre-competitive collaboration to reduce the barriers to innovation. The Ontologies Mapping Project started in 2016 to enable better tools and services for ontology mapping and to define best practices for ontology management in the Life Sciences [1].

The interest in ontologies is growing within the pharmaceutical domain. Data is a very valuable corporate asset to enable digital transformation and lead to innovative biological insight. However, data integration is fundamental piece in the puzzle where ontologies and ontology matching may play an important role.

The Pistoia Alliance Ontologies Mapping Project has covered two domains of interest: *(i)* phenotype and disease [2], and *(ii)* laboratory analytics domain. In this paper we focus on the later, for which alignment sets are not that common, we introduce the system Paxo, and we compare its results against participants of the Ontology Alignment Evaluation Initiative (OAEI, `http://oaei.ontologymatching.org/`).

Datasets. We selected, in conjunction with (pharmaceutical) industry partners of the Pistoia Alliance, 9 relevant ontologies to the laboratory analytics domain and 13 ontology pairs to compute their alignment. Table 1 shows the ontologies that were selected for their relevance to the laboratory analytics domain. Note that there is not a public hand-curated gold standard alignment among the selected ontology pairs.

Paxo system. Paxo is a lightweight ontology mapping approach. Unlike other algorithms, Paxo does not need to store, load or index ontologies. Instead Paxo accesses the API of the Ontology Lookup Service (OLS, `https://www.ebi.ac.uk/ols/index`) and the Ontology Mapping Repository (OxO, `https://www.ebi.ac.uk/spot/oxo/`) at EMBL-EBI to explore ontologies. Through OLS, Paxo can perform search via preferred label and synonyms, while OxO offers access to a wide range of known ontology mappings, that were defined, for example, as cross references within the ontologies themselves or in the UMLS Metathesaurus.

2 Evaluation

Table 2 shows the number of computed mappings, for the 13 selected matching tasks, by Paxo (with relaxed-R and strict-S variants) and a subset of the OAEI systems that were able to cope with (most of) the selected matching tasks.

We have computed consensus alignments of vote 2, 3 and 4 (*i.e.*, mappings suggested by at least two, three or four systems, respectively). Note that, when there are several systems of the same family (*i.e.*, systems participating with several variants), their (voted) mappings are only counted once in order to reduce bias.

Domain	Ontology Name	Acronym	Size	Version
Chemistry	Allotrope Merged Ontology Suite	AFO	1,868	2019/05/10
	Chemical Methods Ontology	CHMO	3,130	2014-11-20
Biology	Ontology for Biomedical Investigations	OBI	3,959	2019-11-12
	Eagle-I Research Resource Ontology	ERO	4,334	23-07-2019
	Mass Spectrometry Ontology	MS	6,855	19:11:2019
	BioAssay Ontology	BAO	7,512	2.5.1
	Experimentals Factors Ontology	EFO	26,510	3.12.0
General	National Cancer Institute Thesaurus	NCIT	154,108	19.11d
	Medical Subject Headings	MESH	539,242	2019ab

Table 1: Ontologies relevant to the laboratory analytics domain.

Matching Task	System mappings						Consensus mappings			
	Paxo-R	Paxo-S	AML	BioPortal	LogMap	LogMapBio	#SF	Con-2	Con-3	Con-4
AFO-CHMO	234	199	214	160	240	247	6	220	200	176
AFO-MESH	149	76	130	39	152	153	4	120	57	32
AFO-NCIT	461	313	361	213	297	315	4	403	224	159
BAO-MESH	273	176	248	112	313	317	4	251	142	81
BAO-NCIT	564	418	249	230	232	250	6	304	255	242
CHMO-MESH	435	222	240	70	252	257	4	229	124	62
CHMO-NCIT	605	343	196	125	171	209	7	215	151	128
EFO-MESH	3,710	2,953	3,392	1,250	3,054	3,344	4	3,140	2,538	1,170
EFO-NCIT	4,297	3,559	(-)	2,442	3,448	4,047	4	3,054	2,477	2,266
ERO-MESH	277	176	165	74	206	205	4	174	120	65
ERO-NCIT	511	343	174	168	168	194	7	234	191	177
MS-NCIT	268	143	73	86	56	57	5	107	86	74
OBI-NCIT	504	302	137	147	142	155	7	186	155	149

Table 2: Number of mappings for the selected matching tasks. (-): a system failed to compute mappings. #SF: number of system families contributing to the consensus. Con-x: consensus mappings with 'x' votes. We focus on the entities defined in the input ontologies and thus ignore entities imported/reused from external ontologies.

Paxo-R is the system that, on average, predicts the highest amount of mappings followed by LogMap-Bio and Paxo-S; while BioPortal includes, on average, the smallest amount. Figure 1 shows a two-dimensional representation of the *Jaccard distances* among the alignments between EFO and MESH. Paxo-R and Paxo-S produce relatively similar mapping sets (as for LogMap and LogMap-Bio). Being close to a consensus mapping set is not necessarily positive; but it means that the computed mappings are similar to the agreement. For example, BioPortal mappings are typically small in size and close to Con-4. Paxo mappings are different from the other system computed mappings. A more detailed (manual) analysis will be conducted in the near future to evaluate the quality of the reported mapping sets.

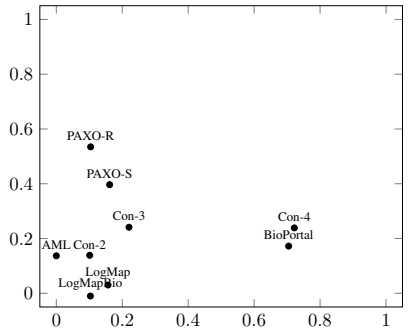

Fig. 1: Two-dimensional representation of the Jaccard distances among EFO-MESH mappings. Plots computed with the MELT framework (`https://github.com/dwslab/melt`).

References

1. Harrow, I., et al.: Ontology mapping for semantically enabled applications. Drug Discovery Today (2019)
2. Harrow, I., Jiménez-Ruiz, E., et al.: Matching disease and phenotype ontologies in the ontology alignment evaluation initiative. J. Biomedical Semantics **8**(1) (2017)