YAM-BIO – Results for OAEI 2017

Amina Annane,[1,2] Zohra Bellahsene, and[1] Faical Azouaou,[2]
Clement Jonquet[1,3]

[1] Laboratory of Informatics, Robotics and Microelectronics of Montpellier (LIRMM)
University of Montpellier & CNRS, France
{annane,jonquet,bella}@lirmm.fr
[2] National Higher School of Informatics (ESI), Algiers, Algeria
{f_azouaou}@esi.dz
[3] Center for BioMedical Informatics Research (BMIR), Stanford University, USA

Abstract. The YAM-BIO ontology alignment system is an extension of YAM++ but dedicated to aligning biomedical ontologies. YAM++ has successfully participated in several editions of the Ontology Alignment Evaluation Initiative (OAEI) between 2011 and 2013, but this is the first participation of YAM-BIO. The biomedical extension includes a new component that uses existing mappings between multiple biomedical ontologies as background knowledge. In this short system paper, we present YAM-BIO's workflow and the results obtained in the *Anatomy* and *Large Biomedical Ontologies* tracks of the OAEI 2017 campaign.

1 Presentation of the YAM-BIO system

1.1 State, purpose, general statement

YAM-BIO may be seen as an extension of YAM++ [5] that uses existing mappings between multiple biomedical ontologies as background knowledge to enhance the matching results. The latest version of YAM++, which we reused in YAM-BIO, obtained excellent results in multiple Ontology Alignment Evaluation Initiative (OAEI) campaigns, especially in 2013 [11]. YAM++ did not participate more since then. Four years on from the last participation, our objective this year was to establish a comparison between the potential performance of a bio-customized YAM++, and state-of-the-art systems in matching biomedical ontologies.

Over last OAEI campaigns, state-of-the-art systems such as AML [7] and LogMapBio [9] used specialized background knowledge to improve their results. More generally, the use of background knowledge –or indirect matching techniques– as recently allowed to obtain better results. YAM-BIO is an equivalent evolution of YAM++ in which we added a component that uses existing mappings as background knowledge. With YAM-BIO, we participated this year to the *Anatomy* and *Large Biomedical Ontologies (Largebio)* tracks.

1.2 YAM-BIO's general alignment worklfow

As illustrated in Fig. 1, YAM-BIO's workflow contains three main steps: First, to compute direct matching between source and target ontologies using YAM++.

Second, to compose relevant existing mappings in the background knowledge for concepts not aligned during first step. Third, to compute union of the alignments produced by the two previous steps.

Direct matching with YAM++: Annotations (labels, comments, etc.) and structures of source and target ontologies are indexed as well as the context of each entity that may be a concept or a property. Then, candidate mappings with a low annotation similarity are pre-filtered. Other advanced lexical and structural similarity measures are applied on the remaining candidate mappings, before updating their similarity scores using the structure information of source and target ontologies. Finally, a threshold is dynamically computed to select the most relevant mapping candidates. For more details on each steps of the execution of YAM++, readers may refer to [5].

Indirect matching and union: During this step YAM-BIO finds mappings for the concepts that have not been matched during direct matching with YAM++. First, background knowledge existing mappings are loaded in a list of lists noted A as follows:

1. Identifiers of all concepts in the background knowledge are added to A. The identifier of a given concept is the last part of its URI, for example the identifier of the concept that has the URI http://mouse.owl#MA_0000031 is MA_0000031.
2. Each element x of A points to a list that contains identifiers of all concepts matched to x in the background knowledge.

Then, for each source concept y that is not matched yet, YAM-BIO checks if y's identifier exists in A. If yes, YAM-BIO gets the corresponding list –pointed by y– and for each element of this list, YAM-BIO verifies if itself points to a list that contains a concept identifier from the target ontology. If so, YAM-BIO derives a new mapping and adds it to the alignment produced previously by the direct matching.

1.3 Adaptations made for the OAEI campaign

The existing mappings used as background knowledge have been extracted from Uberon [10] and the Human Disease Ontology (DOID) [13]. These ontologies contain several manually edited/curated cross references to other biomedical ontologies that we may consider as mappings.

In addition, concept identifiers of the ontologies provided for the Largebio track are not the original ones, but have been replaced by their standardized preferred labels. For this reason, we have used the NCBO BioPortal's REST API [6] to replace concept identifiers within Uberon and DOID by their standardized preferred labels.

1.4 Availability

YAM++ has now a publicly accessible online prototype version [16] and is registered on Maven repositories: http://yamplusplus.lirmm.fr. YAM-BIO has not been packaged yet to be reused by others. However, the alignment set produced

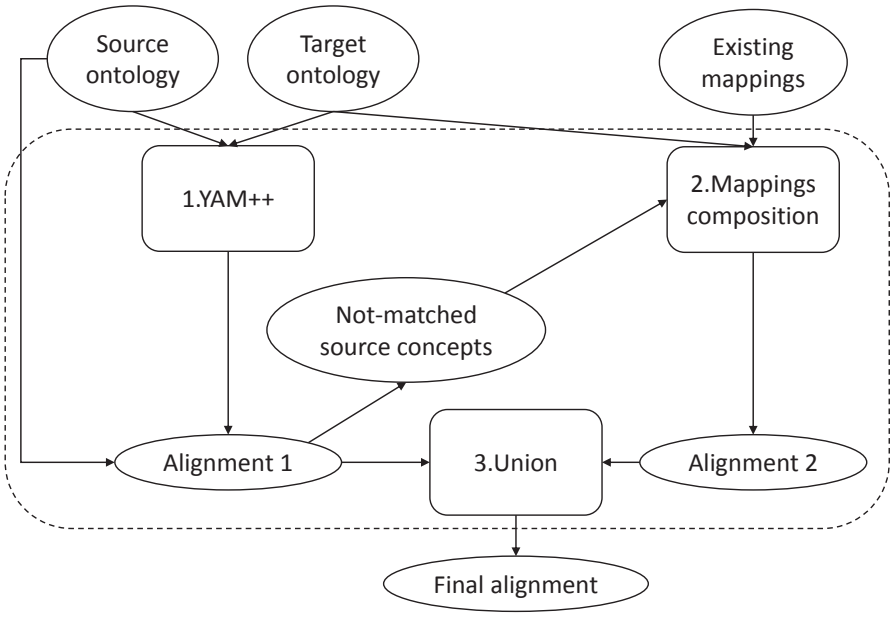

Fig. 1. YAM-BIO's general workflow

as well as the background knowledge file are available at the following link: https://goo.gl/zNznNz

2 Results

2.1 Anatomy track

The Anatomy track consists of finding an alignment between the Adult Mouse Anatomy [8] (2744 classes) and a subset of the National Cancer Institute (NCI) Thesaurus [14] (3304 classes) describing human anatomy. Table 1 shows YAM-BIO's evaluation result and runtime on this track. YAM-BIO scored in second position among the 12 systems that have participated in 2017 with almost the same precision and a slightly lower recall comparing to the top ranked system.

Table 1. YAM-BIO's Anatomy track results

Test set	Precision	Recall	F-Score	Time (s)
Anatomy	0.948	0.922	0.935	70

2.2 Large Biomedical Ontologies (Largebio) track

The Largebio track consists of respective finding alignments between the Foundational Model of Anatomy (FMA) [12], SNOMED-CT [4], and the NCI Thesaurus. There are six tasks with different input ontology sizes: small fragment, large fragment and whole ontologies. Table 2 shows YAM-BIO's evaluation re-

sults and runtime on those tasks. With the exception of the XMAP system[4], YAM-BIO is the top ranked system in Task 1 and Task 4 and obtained almost the same results as the best system in Task 3 with an F-measure of 0.834 vs 0.835. In Task 2 and Task 6, YAM-BIO scored in second position with a better recall than the best system and a lower precision. In Task 5, it shared third position with LogMapBio. In terms of running time, YAM-BIO completed the different tasks in acceptable time.

Table 2. YAM-BIO's LargeBio track results

Test set	Precision	Recall	F-Score	Time (s)
Task 1: Small fragments FMA-NCI	0.968	0.896	0.931	56
Task 2: FMA Whole-NCI Whole	0.816	0.888	0.850	279
Task 3: Small fragments FMA-SNOMED	0.966	0.733	0.834	60
Task 4: FMA Whole-SNOMED Large fragment	0.887	0.728	0.800	468
Task 5: Small fragments SNOMED-NCI	0.899	0.677	0.772	2202
Task 6: SNOMED Large fragment-NCI Whole	0.827	0.698	0.757	490

3 Discussion

3.1 Comments on the results and ways of improvement

YAM-BIO scored second position in the Anatomy track and scored first or second also in the Largebio track (except Task 5). As expected, using existing mappings as background knowledge has improved YAM++ results in terms of recall and consequently F-measure. Mapping compositions extracted from Uberon allowed YAM-BIO to discover non trivial mappings, specifically in Anatomy track and in Task 1 and Task 2 of Largebio track. Similarly, the composition of mappings extracted from DOID allowed to increase the recall of Task 5 and Task 6. However, the incoherence analysis shows that YAM-BIO returns some incoherent mappings. This may be explained by the fact that the mappings derived using background knowledge have been added to the final alignment without any semantic verification.

In our current system, mappings derived using background knowledge are not post-filtered and semantically verified as in YAM++. A simple union of the direct and indirect alignments is performed to obtain the final alignment. In the future, our goal would be to integrate the use of background knowledge directly inside YAM++'s internal architecture which, we believe, will improve coherence of the final results. More specifically, we will implement the approach proposed in [1].

In addition, we are aware of the importance of the dynamic selection of ontologies to use as background knowledge [15, 2]. Indeed, from the selected ontologies we may extract manual/automatic mappings as background knowledge. For this reason, we will extend YAM-BIO to dynamically select a set of ontolo-

[4] We note XMAP uses UMLS Metathesaurus as background knowledge, which is the same from which Largebio reference alignments are extracted.

gies from a given ontology library such as the NCBO BioPortal or Watson [3], if we want to go beyond biomedicine.

3.2 Comments on the OAEI evaluation

When possible, we think it would be interesting to publish participants results with and without use of specialized background knowledge. On one hand, this will allow to better evaluate the influence of background knowledge in matching quality and running time. On the other hand, this will allow a fair comparison with systems that do not use background knowledge.

Some components are common in all ontology matching system architectures; others do not always exist —such as background knowledge selection or semantic verification. This makes the comparison of running time executions particularly cumbersome and not always fair. According to us, it would be more appropriate to evaluate execution times for each separate component. For example, YAM-BIO used a predefined background knowledge while LogMapBio made a dynamic selection from an online repository necessarily taking additional time. Splitting running time by components will also help the community to identify less efficient components to improve them, and most efficient ones to reuse them.

4 Conclusion

In 2017 YAM-BIO participated in two tracks: Anatomy and LargeBio. The results obtained in those tracks are very close to top ranked state-of-the-art systems, thanks to different content matching techniques implemented in YAM++ and to the use of background knowledge. Due to the high heterogeneity of ontologies, we believe that an advanced generic (i.e., not restricted to biomedicine) module that selects and uses background knowledge should be implemented in the internal architecture of YAM++ to improve its results. In the future, we will work on such a module and hopefully participate in different OAEI tracks.

5 Acknowledgment

This work was done during a LIRMM-ESI collaboration within the Semantic Indexing of French biomedical Resources (grant ANR-12-JS02-01001) and PractikPharma (ANR-15-CE23-0028) projects that received funding from the French National Research Agency as well as by the European H2020 Marie Sklodowska-Curie action (agreement No 701771), the University of Montpellier and the CNRS. The authors also acknowledge the Eiffel Excellence Scholarship program.

References

1. Annane Amina, Bellahsene Zohra, Azouaou Faical, and Jonquet Clement. Selection and combination of heterogeneous mappings to enhance biomedical ontology matching. In *20th International Conference on Knowledge Engineering and Knowledge Management, EKAW, Bologna, Italy*, pages 19–33, 2016.
2. Faria Daniel, Pesquita Catia, Santos Emanuel, Cruz Isabel F, and Couto Francisco M. Automatic background knowledge selection for matching biomedical ontologies. *PloS One*, 9(11):e111226, 2014.
3. d'Aquin Mathieu, Gridinoc Laurian, Angeletou Sofia, Sabou Marta, and Motta Enrico. Watson: A Gateway for Next Generation Semantic Web Applications. In

6th International Semantic Web Conference, ISWC, Poster and Demonstration, Busan, Korea, pages 11–15, 2007.

4. Kevin Donnelly. SNOMED-CT: The advanced terminology and coding system for eHealth. *Studies in health technology and informatics*, 121:279, 2006.
5. Ngo DuyHoa and Bellahsene Zohra. Overview of YAM++:(not) yet another matcher for ontology alignment task. *Journal of Web Semantics*, 41:30 – 49, 2016.
6. Noy Natalya F, Shah Nigam H, Whetzel Patricia L, Dai Benjamin, Dorf Michael, Griffith Nicholas, Jonquet Clement, Rubin Daniel L, Storey Margaret-Anne, and Chute Christopher G. BioPortal: ontologies and integrated data resources at the click of a mouse. *Nucleic Acids Research*, 37:170–173, 2009.
7. Daniel Faria, Catia Pesquita, Booma S Balasubramani, Catarina Martins, Joao Cardoso, Hugo Curado, Francisco M Couto, and Isabel F Cruz. Oaei 2016 results of AML. In *11th International Workshop on Ontology Matching, Kobe, Japan.*, pages 138–145, 2016.
8. Terry F. Hayamizu, Mary Mangan, John P. Corradi, James A. Kadin, and Martin Ringwald. The adult mouse anatomical dictionary: a tool for annotating and integrating data. *Genome Biology*, 6(3):R29, Feb 2005.
9. E Jiménez-Ruiz, B Cuenca Grau, and V Cross. Logmap family participation in the oaei 2016. In *11th International Workshop on Ontology Matching, Kobe, Japan.*, pages 185–189, 2016.
10. Christopher J. Mungall, Carlo Torniai, Georgios V. Gkoutos, Suzanna E. Lewis, and Melissa A. Haendel. Uberon, an integrative multi-species anatomy ontology. *Genome Biology*, 13(1):R5, Jan 2012.
11. DuyHoa Ngo and Zohra Bellahsene. YAM++ results for OAEI 2013. In *8th International Workshop on Ontology Matching, Sydney, Australia.*, pages 211–218, 2013.
12. Cornelius Rosse and Jos L.V. Mejino. A reference ontology for biomedical informatics: the foundational model of anatomy. *Journal of Biomedical Informatics*, 36(6):478 – 500, 2003. Unified Medical Language System.
13. Lynn Marie Schriml, Cesar Arze, Suvarna Nadendla, Yu-Wei Wayne Chang, Mark Mazaitis, Victor Felix, Gang Feng, and Warren Alden Kibbe. Disease ontology: a backbone for disease semantic integration. *Nucleic Acids Research*, 40(D1):D940–D946, 2012.
14. Nicholas Sioutos, Sherri de Coronado, Margaret W. Haber, Frank W. Hartel, Wen-Ling Shaiu, and Lawrence W. Wright. NCI thesaurus: A semantic model integrating cancer-related clinical and molecular information. *Journal of Biomedical Informatics*, 40(1):30 – 43, 2007. Bio*Medical Informatics.
15. Chen Xi, Xia Weiguo, Jiménez-Ruiz Ernesto, and Cross Valerie. Extending an ontology alignment system with BioPortal: a preliminary analysis. In *13th International Semantic Web Conference, ISWC, Posters and Demonstrations, Riva del Garda, Italy*, pages 313–316, 2014.
16. Bellahsene Zohra, Emonet Vincent, Ngo DuyHoa, and Todorov Konstantin. Yam++ online: a multi-task platform for ontology and thesaurus matching. In *14th Extended Semantic Web Conference, ESWC, Posters and Demonstrations, Portoroz, Slovenia*, 2017.