Legato: Results for OAEI 2017

Manel Achichi, Zohra Bellahsene, Konstantin Todorov

{firstname.lastname}@lirmm.fr
LIRMM / University of Montpellier, France

Abstract. *Legato* is an automatic data linking system handling datasets containing blocks of highly similar in their descriptions but yet distinct resources, as well as resources with highly heterogeneous descriptions. This paper presents the results of *Legato* on the Instance Matching track of the Ontology Alignment Evaluation Initiative 2017 via the SEALS platforme. *Legato* participated in the two sub-tracks of the instance matching track. We briefly describe the *Legato* framework, we present the different techniques used by the system in the accomplishment of the data linking task and we present and discuss the alignment results of the system as compared to the other tools participating to the 2017-edition of the evaluation campaign.

1 Presentation of the System

We begin by providing an overview of the main characteristics of *Legato*, as well as describing briefly the specific techniques applied in the different parts of its workflow.

1.1 General Features and Purpose

Legato is a data linking tool developed in the framework of the DOREMUS project[1]. It is designed to match entities from highly heterogeneous graphs, effectively disambiguating highly similar (yet distinct) resources. *Legato* is based on indexing techniques, with a preliminary phase of data cleaning allowing to prune properties that make the comparison task difficult, as well as a post-processing phase allowing to discard erroneous links and to lower the rate of false positives. An important feature of our system is that it requires very little manual configuration – neither similarity measures and thresholds, nor properties to align are required as input. The values of the various thresholds inherent to the algorithm are set empirically so as to ensure a maximum performance on a large variety of heterogeneous data. With this, we aim at placing *Legato* among the few fully automatic instance matchers in the state of the art. The system is openly available at the following link: `https://github.com/DOREMUS-ANR/legato`.

[1] `http://www.doremus.org/`

1.2 Specific Techniques Used

This section briefly describes the overall workflow of *Legato*, shown in Figure 1. Its configuration takes one single parameter: the type of resources for comparing and linking. The system then proceeds to automatically process, compare, repair and provide a set of identity links (`owl:sameAs` statements). More precisely, *Legato* implements the following successive steps.

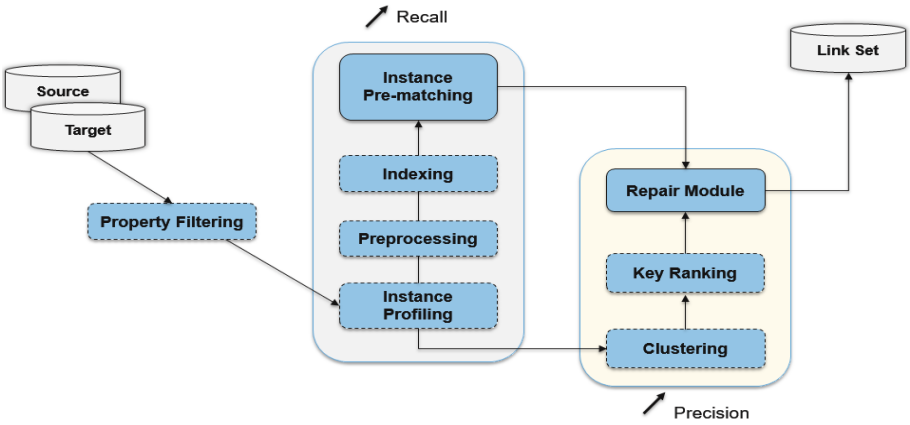

Fig. 1: The processing pipeline of *Legato*

Data cleaning. The first step before representing the resources in a comparable form consists in filtering the problematic properties from the two input datasets. *Legato* considers a property as *problematic* if it hinders the comparison of resources. Consider the example given in Table 1, issued from the DOREMUS track data from the IM@OAEI2017 (Instance Matching track of the Ontology Alignment Evaluation Initiative).

The descriptions `mw1` and `mw1'` are about two equivalent musical works retrieved from *Philharmonie de Paris* (PP) and *Bibliothèque Nationale de France* (BNF), respectively. These descriptions are highly similar, with the notable exception of the respective `ecrm:P3_has_note` property values. Considering this property, we would yield a very low value of the similarity score, and still it is likely that this property is discovered as a key (because of its unique values) and therefore used in a configuration file of a linking system.

Properties identified as problematic may concern those that have values in a free text format, i.e., comments (as in the example above), as well as resource-specific values, that the publisher cannot describe freely. For example, for the same musical work, two institutions would generally assign different identifiers in their respective catalogs. The way we propose to identify automatically problematic properties, is to discover mono-property keys valid **on both** datasets, i.e., each object for such a property has at most one subject in both datasets.

`mw1`[2]	a efrbroo:F22_SelfContained_Expression mus:U70_has_title "Sonates" mus:U12_has_genre sonate[3] ecrm:P3_has_note "Cette sonate est constituée de cinq formants: Antiphonie, Trope, Constellation, Strophe et Séquence. Seuls les 2e et 3e formants sont publiés. Le Formant 2 (Trope) est composé de quatre sections : Commentaire, Glose , Texte, Parenthèse, qui peuvent être jouées dans différents ordres. Cette oeuvre nécessite un piano à 3 pédales. - Durée d'exécution : 20 minutes environ"
`mw1'`[4]	a efrbroo:F22_SelfContained_Expression mus:U70_has_title "Sonates" mus:U12_has_genre sonate[5] ecrm:P3_has_note "Date de révision : 1963, comprend : Antiphonie; Trope; Constellation (ou Constellation-Miroir); Strophe; Séquence"

Table 1: `ecrm:P3_has_note` — An example of a problematic property in DOREMUS data

Instance profiling. *Legato* creates instance profiles by exploiting the information in the $\mathcal{CBD}$s (for Concise Bounded Description) of the resources.[6] We extend the $\mathcal{CBD}$ notion by also considering the descriptions of neighboring nodes of a resource in its graph. At this step, *Legato* extracts a subgraph for each resource `r` that includes all the triples from the $\mathcal{CBD}$ of `r`, the $\mathcal{CBD}$s of its direct predecessors (linked by incoming links to `r`), and the $\mathcal{CBD}$s of its direct successors (linked through outgoing links to `r`). For instance profiling, *Legato* only considers datatype properties. In that, each resource is represented by a set of literals in its profile (subgraph) considered as relevant for its description. This strategy allows to avoid manually setting the graph traversal distance to which the information should be collected.

Instance pre-matching. Once all resources in both datasets are profiled, *Legato* employs an indexing technique to project each profile onto a vector space where terms are weighted by their TF-IDF (Term Frequency-Inverse Document Frequency) values. Two standard NLP (Natural Language Processing) filters are applied: tokenization and stop-words removal. Finally, *Legato* pre-selects the identity links by computing the correlation between vectors by using the well-known cosine similarity. In order to increase recall and to automate the threshold setting independently on the data, at this stage *Legato* generates links with a very low threshold (empirically fixed at 0.2).

Link repairing. To ensure coherence, the alignments selected at the pre-matching step are passed to the *repair* module. Note that decreasing the similarity threshold may increase the number of false positive matches. As indicated above, a source resource may be erroneously aligned to many target resources

[6] `https://www.w3.org/Submission/CBD/`

(and vice versa). This is due to the fact that we can have highly similar descriptions of different resources in a single dataset. Therefore, *Legato* includes a post-processing phase allowing to disambiguate between such resources and to repair the erroneous links generated between them in the previous phase. We employ a clustering algorithm [1] within each dataset aiming to group together the similar resources. Then, for each pair of similar clusters (identified by a cluster matching algorithm) across the two datasets, the resources are compared on a best-key basis. We apply the RANKey algorithm for identifying and ranking the key properties [2]. For each link $\mathtt{l}=(\mathtt{r_s}, \mathtt{r_t})$ produced in the earlier step, the repair module begins by searching for a link of $\mathtt{r_s}$ to a target resource $\mathtt{r'_t} \neq \mathtt{r_t}$, based on the key strategy. If found, the target resource $\mathtt{r_t}$ in $\mathtt{l}$ is then replaced by $\mathtt{r'_t}$. In case multiple matches are found in that scenario, the one with the highest similarity score is kept. The repair module aims at improving precision.

Link to the System and Parameters File. We provide an open source implementation of *Legato* in a GitHub project under the following link: `https://github.com/DOREMUS-ANR/legato`. It is available as an eclipse project. *Legato* provides an appropriate user interface allowing the user to select the source, target and alignment (if it is available) files for aligning and evaluating the produced links. If no alignment file exists, *Legato* produces a set of identity links without evaluating them.

Link to the Set of Provided Alignments. The alignments produced by *Legato* on the instance matching track of OAEI2017 can be downloaded at `https://github.com/manoach/Legato-at-OAEI-2017`.

2 Results

In this section, we present the results obtained by *Legato* on the data coming from the instance matching track of the OAEI2017 campaign.[7] This year, the instance matching track contains two tasks and four datasets. *Legato* participated to all these tasks.

2.1 Synthetic Task

This task contains synthetic data about creative works. They have been generated through the Semantic Publishing Instance Matching Benchmark (SPIMBENCH) [3] by transforming the source instances based on their values, structure and semantics. The task contains two matching sub-tasks on two different datasets: SPIMBENCH sandbox and SPIMBENCH mainbox (datasets of different sizes). The first one contains 380 resources while the second one – 1800.

Tables 2 and 3 show *Legato*'s results as compared to those of the other systems that have participated at this task, namely, AML, I-Match and LogMap. As it

[7] `http://oaei.ontologymatching.org/2017/`

System	Precision	Recall	F-measure
AML	0.849	**1.000**	0.918
I-Match	0.854	0.997	**0.920**
Legato	**0.980**	0.730	0.840
LogMap	0.938	0.763	0.841

Table 2: Results for SPIMBENCH sandbox.

System	Precision	Recall	F-measure
AML	0.855	**1.000**	**0.922**
I-Match	0.856	0.997	0.921
Legato	**0.970**	0.700	0.810
LogMap	0.893	0.709	0.790

Table 3: Results for SPIMBENCH mainbox.

can be seen, *Legato* achieves the highest score in terms of precision for both SPIMBENCH sandbox and SPIMBENCH mainbox (98% and 97%, respectively). We notice that Legato performs overall well on this task achieving a recall of 73% and 70%, and F-measures of 84% and 81% for SPIMBENCH sandbox and SPIMBENCH mainbox, respectively.

2.2 DOREMUS Task

The data from the DOREMUS track contain descriptions of real-world classical music works and events, coming from the catalogs of two major French cultural institutions (the Philharmonie de Paris and the National Library). These data have been converted to RDF from their original MARC format by the help the specifically designed for that purpose by the DOREMUS team tool *marc2rdf*.[8] These data follow a common ontology [4] given by the DOREMUS model, extending well-established models for intellectual works description, historically used by libraries.[9]

System	Precision	Recall	F-measure
AML	0.851	0.479	0.613
I-Match	0.680	0.071	0.129
Legato	0.930	0.920	0.930
LogMap	0.406	0.882	0.556
NjuLink	**0.966**	**0.945**	**0.955**

Table 4: Results for HT of the DOREMUS task

System	Precision	Recall	F-measure
AML	0.914	0.427	0.582
I-Match	**1.000**	0.053	0.101
Legato	**1.000**	**0.980**	**0.990**
LogMap	0.119	0.880	0.210
NjuLink	0.959	0.933	0.946

Table 5: Results for FPT of the DOREMUS task

Tables 4 and 5 show *Legato*'s results and those of the four other systems that participated at this task, namely, AML, I-Match, LogMap and NjuLink.

[8] `https://github.com/DOREMUS-ANR/marc2rdf`
[9] `http://data.doremus.org/ontology/`

On both subtasks, two systems stand out in terms of performance – *Legato* and NjuLink, achieving comparable results and outperforming considerably the other participant systems. More precisely, on the Heterogeneities task (HT data), *Legato* ranks second after NjuLink with a precision of 93%, a recall of 92% and F-measure of 93%. As for the False Positives Trap task (FTP data), it can be seen in Table 5 that *Legato* achieves the best results in terms of precision (100%), recall (98%) and F-measure (99%). It is worth noting that the DOREMUS track appeared to be problematic for the majority of the systems, with average F-measure scores of around 0.6 over all participants on both tasks.

3 Discussion

As seen in the previous section, our system proves to be very effective for the two sub-tracks of the instance matching track of OAEI 2017, showing its strength of producing high scores in terms of F-measure (above 80% on all tasks). *Legato* produced the best precision in 3 of the 4 instance matching tasks. Thanks to its repair module, *Legato* ensures a very high accuracy, which is no less than 93% on all instance matching tasks. In terms of recall, *Legato* scored well on the DOREMUS track, but obtained the lowest rank on the synthetic data track. We explain that result by the fact that *Legato* does not yet tackle value-based variations that are characteristic for the synthetic data – the lack of lemmatization in the indexing process of our system equates to looking only for exact matches between string values.

Proposed Improvements of the System *Legato* implements an approach handling structurally heterogeneous descriptions. However, the limit of the current version of our system is that it is not dealing with value-based heterogeneity, but rather considers exact matches only. Therefore, this will be the main base of future improvements. Furthermore, we plan to discover matches between resources coming from multiple data sources simultaneously.

4 Conclusion

In this paper, we presented *Legato*—an automatic and generic data linking tool. *Legato* participates for the first time at the OAEI campaign and it was evaluated on data from the two sub-tracks of the Instance Matching track. The results showed that *Legato* is capable of effectively linking both synthetic and real-world data of highly heterogeneous nature achieving comparable results to the best systems and outperforming most of them in terms of precision while keeping a decent recall level. In addition, *Legato* achieved the best score on the FPT DOREMUS data containing highly similar resources, thanks to its post-processing link repairing step. Finally, *Legato* is among the few participant systems that are freely available and ready to use by researchers or practitioners.

Acknowledgements

This work has been partially supported by the French National Research Agency(ANR) within the DOREMUS Project, under grant number ANR-14-CE24-0020.

References

1. L. Rokach and O. Maimon, "Clustering methods," in *The Data Mining and Knowledge Discovery Handbook.*, pp. 321–352, 2005.
2. M. Achichi, M. Ben Ellefi, D. Symeonidou, and K. Todorov, "Automatic key selection for data linking," in *Knowledge Engineering and Knowledge Management: 20th International Conference, EKAW 2016, Bologna, Italy, November 19-23, 2016, Proceedings 20*, pp. 3–18, Springer, 2016.
3. T. Saveta, E. Daskalaki, G. Flouris, I. Fundulaki, M. Herschel, and A.-C. Ngonga Ngomo, "Pushing the limits of instance matching systems: A semantics-aware benchmark for linked data," in *Proceedings of the 24th International Conference on World Wide Web*, pp. 105–106, ACM, 2015.
4. M. Achichi, R. Bailly, C. Cecconi, M. Destandau, K. Todorov, and R. Troncy, "Doremus: Doing reusable musical data," in *ISWC: International Semantic Web Conference*, 2015.