# ILP for Mining Linked Open Data: a biomedical Case Study

Gabin Personeni, Simon Daget, Céline Bonnet, Philippe Jonveaux,
Marie-Dominique Devignes, Malika Smaïl-Tabbone, Adrien Coulet

# ILP for Mining Linked Open Data: a biomedical Case Study

Gabin Personeni[1], Simon Daget[1], Céline Bonnet[2], Philippe Jonveaux[2], Marie-Dominique Devignes[1], Malika Smaïl-Tabbone[1], and Adrien Coulet[1]

[1] LORIA (CNRS, Inria NGE, Université de Lorraine)
[2] Laboratoire de Génétique Médicale, CHU de Nancy, INSERM U-954

**Abstract.** This paper is a summary of a recent work accepted (as a long paper) for publication in the tenth international conference on Data Integration in Life Sciences (DILS 2014) [1]. We briefly describe in the last section ongoing work for improving the rule-based prediction process.
Increasing amounts of biomedical data provided as Linked Open Data (LOD) offer novel opportunities for knowledge discovery. LOD are represented in a standard format, partially integrated, and offer connections with domain knowledge available in semantic web ontologies. Relational data mining methods such as ILP are good candidates to consider together LOD and domain knowledge awareness. We propose in this paper an approach for collecting and mining LOD, using ILP, with the goal of characterizing and predicting genes responsible for a disease. An integration step enables to select and link together relevant pieces of LOD. It results from this integration a graph that is subsequently mined using ILP. For this real-world use case, we design ILP experiments on two subsets of relational descriptors of genes responsible for intellectual disability. We evaluate ILP results and assess the contribution of domain knowledge. Our ongoing efforts explore how the combination of rules coming from distinct theories can improve the prediction accuracy.

## 1 Introduction

Linked Open Data (LOD) are part of a community effort to build a semantic web, where web resources can be interpreted both by humans and machines. LOD are available as a large and growing collection of datasets represented in the standard RDF format, partially connected to each other and to domain knowledge represented within semantic web ontologies [2].

This recent availability of LOD can be particularly beneficial to the life sciences, where relevant data are spread over various data resources with no agreement on a unique representation of biological entities [3]. Consequently, data integration is an initial challenge one faces for mining life science data. Various initiatives such as Bio2RDF or the EBI platform aim at pushing life sciences data into the LOD with the will of facilitating their integration [4, 5]. Furthermore, LOD may be connected to domain knowledge represented within ontologies such as the Gene Ontology [6]. Ontologies provide a formal representation of a particular domain that can be used to support automatic reasoning. We have investigated that ontologies and their associated reasoning mechanisms can be

coupled with data mining to facilitate the process of knowledge discovery [7]. We would like to extend this investigation to the context of LOD.

We propose here an approach to: *(1)* select LOD matching a conceptualization of data related to a biomedical question, *(2)* integrate them using existing or newly generated links and then *(3)* mine LOD with Inductive Logic Programming (ILP) using domain knowledge defined in ontologies. The next section presents a state of the art of data mining applied to LOD. The third section presents the LOD selection and integration. The fourth section reports about ILP experiments on selected LOD. The last section describes ongoing experiments.

## 2 State of the Art

The complexity of LOD has motivated several studies about the preparation of data before mining. Callahan *et al.* proposed to map LOD from various datasets to an upper-level ontology named SIO [8]. This ontology serves consequently as a global schema and its terms are used to write federated queries over LOD datasets. SADI is a general framework to facilitate the discovery and use of web services [9]. Because it has been developed with semantic web technologies, SADI is well adapted to define pipelines that can query SPARQL endpoints and integrate their results. The COEUS platform follows a similar rationale but includes a federation layer that facilitates data integration [10]. Any of these solutions is well adapted when either entities have a unique URI over distinct datasets, or when links have been defined between datasets. Unfortunately, these two prerequisites are not guaranteed in LOD [11, 12]. In this work, we want to use LOD datasets that have been developped independently, even if this requires the definition of novel mappings between datasets. For this reason we propose a simple but generic way for selecting and integrating LOD before their mining.

The emergence of workshops about the mining of LOD illustrates the rise of interest for this topic, both in the semantic web and data mining communities. Some contributions in this domain aim at completing or correcting the LOD. For instance, Gangemi *et al.* proposed an approach to type systematically DBpedia entities using graph patterns and disambiguation techniques [13]. Other authors studied how to propose systematically missing links, particularly between unrelated datasets [14, 15]. For example, a tool for defining `owl:sameAs` links in the LOD between equivalent drugs has been proposed by Brenninkmeijer *et al.* [16].

A second group of works explores how some peculiarities of LOD can help data mining. For example, Percha *et al.* used paths between distinct drugs in linked data to predict novel drug-drug interactions [17]. Here, the fact that relationships and entities are typed in LOD enable to define features that characterise possible paths between drugs and consequently to train a random forest classifier. Pathak *et al.* proposed a study on how federated queries over Electronic Health Records and drug related LOD could enable the discovery of novel drug-drug interactions [18].

To our knowledge, only few seminal works have explored how LOD mining can take advantages of knowledge representation [19, 20]. In this work we propose to explore this direction using ILP in a biomedical case study.

## 3 LOD Selection and Integration

The first step, of conceptualization, is to build an entity-relationship (ER) model decribing the entities relevant to consider for a given study. This step is realized with a domain expert, and does not require any knowledge of what data is available in LOD and how it is structured. An ER model consists of a conceptualization usually made of entities, relationships and attributes. We use only a subset of those: entities and binary relationships without attributes (similarly to RDF properties). In our case, $n$-ary relationships and relationships with attributes are represented with a composition of binary relationships using reification.

LOD integration consists primarily in mapping our ER model onto LOD types of entities and of relationships. This mapping is materialized by defining correspondances between each entity of the model and one or many RDF entity types of LOD; and between each relationship of the model and RDF properties present in LOD. Each entity is further defined by a *concept definition* that can be either an RDF entity type, its negation, the domain/range of a property, or the union/intersection of two entity types. Similarly, the relationships of the ER model can be mapped to one property or a composition of properties (or inverse properties).

Because the mapping can associate one entity with several datasets, it can cause redundancy. To guarantee the consistency of collected data, we need additional information on individual identity. Individuals are identified in LOD by their URIs. The main issue in mining LOD from several datasets is that two distinct URIs from different LOD datasets may refer to the same real world object. Equivalence between individuals is expressed using the property `owl:sameAs`, sometimes with the less precise property `rdfs:seeAlso` or with a dataset dependent property. We propose an automatic way for resolving identity of individuals, using equivalence links when available, and creating new one when none exists. New links are created on the basis of individual URIs, related data and LOD structure. Links we generated this way are available at `http://www.loria.fr/~coulet/dils14/individual_identities.html`.

## 4 ILP Mining of Linked Open Data

### 4.1 ILP Experiments

The aim of the mining step is to learn by ILP the concept of genes responsible for Intellectual Disability (ID) from the set of integrated triples relative to positive and negative examples of such genes. The experiments reported in this paper were conducted with the Aleph program [21]. Aleph parameters *rule size, minpos, noise, minacc* have been set respectively to 6, 5, 3 and 85%. The *noise* parameter allows rules to tolerate a few exceptions and constitutes an advantage when dealing with noisy data such as biological LOD.

The outcome of the mining experiment is used both for predictive and descriptive purposes. The predictive power of the first-order logic (FOL) rules is evaluated by cross-validation whereas their descriptive power is analyzed qualitatively.

Our first experiment ($G1$) applies to the genes, their features (*i.e.,* protein they produces, pathways they are involved in, etc.) and their GO annotations plus their direct parent using the *is-a* relationship defined in the Gene Ontology. The aim of the next three experiments $G2$, $G3$ and $G4$ is to assess the contribution of domain knowledge when allowing respectively 2, 3 and 4 generalization steps on the GO structure, which is a rooted directed acyclic graph. For that, we add inference rules expressing the transitivity of the *is-a* relationship and the number of steps to consider. Examination of the resulting 4 theories revealed that the produced rules mostly contain predicates related to GO-terms. Other predicates occur rarely, because GO annotations are plethoric compared to other data. This motivated fifth experiment (named $no-GO$) aimed at analyzing all features excepting the GO-term annotations.

### 4.2 Evaluation of the Results

Complete theories produced in the five experiments are accessible online at `http://www.loria.fr/~coulet/dils14/theories.pdf`. We evaluate the predictive power of each theory using cross-validation. Dedicated KNIME workflows were used for that purpose [22]. A gene is predicted as responsible for ID if it is covered by at least one rule of the theory. Table 1 reports the results of the leave-one-out cross-validation of ILP learning for the experiments $no-GO$ and $G1$ to $G4$. Results show that without GO-term facts ($no-GO$), the prediction accuracy is rather low (59.6%) with a high specificity but a very low sensitivity. Using GO-terms improve prediction indicators: the accuracy increases to 69.8% as we allow Aleph to use more domain knowledge (by performing more generalization). Because a comparative study is hard to adapt to our approach, we propose a qualitative analysis in the next section.

**Table 1.** Results of the leave-one-out cross-validation theories produced by the 5 experiments: True/False Positives, True/False Negatives, Sensitivity, Specificity, Accuracy.

| Experiment | TP | FP | TN | FN | Sens.(%) | Spec.(%) | Acc.(%) |
|---|---|---|---|---|---|---|---|
| $no-GO$ | 75 | **15** | **252** | 207 | 26.6 | **94.4** | 59.6 |
| $G1$ | 135 | 50 | 217 | 147 | 47.9 | 81.3 | 64.1 |
| $G2$ | 157 | 52 | 215 | 125 | 55.7 | 80.5 | 67.8 |
| $G3$ | 157 | 49 | 218 | 125 | 55.7 | 81.7 | 68.3 |
| $G4$ | **161** | 45 | 222 | **121** | **57.1** | 83.1 | **69.8** |

### 4.3 Qualitative Analysis and Discussion

We analyze the theories from a descriptive point of view: in the absence of GO-term facts, we observe several rules containing predicates related to genome location pointing to chromosomes 1 and X, and the chromosomic location 22q13 as possible reservoirs of genes responsible for ID. Other rules point to pathways involved in the metabolism of the cell. Indeed inherited metabolic disorders are considered as an important etiology for ID [23].

In the presence of GO-term facts (experiments $G1$ to $G4$), the repertoire of GO-terms appearing in the rules either as direct protein annotation or as

common ancestor after generalization varies with the experiment and the generalization degree. Such rules suggest that the descriptive power of the theories increases when domain knowledge is taken into account. The value of adding generalization can be illustrated on rules referring to 'organonitrogen compound metabolism'. Rules from $G1$ and $G2$ theories points to the term 'organonitrogen compound **metabolic** process', covering 23 positive examples, while rules from $G3$ and $G4$ theories points to the more specific term 'organonitrogen compound **catabolic** process' covering up to 42 positive examples. Thus allowing for more generalization steps yields more compact theories, an increase in rule coverage and a better specification of features shared by positives examples.

## 5 Ongoing work

The published results, obtained on a biomedical real-world problem show that we successfully integrated Linked Open Data despite missing links. We also confirm that ILP is well fitted for learning in this context as it allowed us to exploit domain knowledge in a unified approach.

We are currently exploring how to design more accurate prediction models from the collected data. The first option underway is to build theories on distinct descriptor subsets and combine the resulting rules as suggested in [24]. We decide to use and compare three propositional classifiers: decision trees, SVM, and neural networks. In these experiments we consider the condition part of ILP rules as both propositional and local gene features. We prefer this strategy to the direct propositionalization of the relational data because we want to minimize information loss [25]. Furthermore we ask Aleph to generalize each positive example (*induce_max* option) in order to get the largest set of local regularities. The second option will consist in applying boosting and bagging ensemble mechanisms to the Aleph program. In parallel, we aim at assessing and comparing the generalization power of the various prediction models by building a larger test set with domain experts.

## References


[1] Gabin Personeni *et al.* Mining Linked Open Data: a Case Study with Genes Responsible for Intellectual Disability. In *DILS*. Springer, 2014.

[2] Christian Bizer *et al.* Linked Data - The Story So Far. *Int. J. Semantic Web Inf. Syst.*, 5(3):1–22, 2009.

[3] Erick Antezana *et al.* Biological knowledge management: the emerging role of the Semantic Web technologies. *Briefings in Bioinformatics*, 10(4):392–407, 2009.

[4] Francois Belleau *et al.* Bio2RDF: Towards a mashup to build bioinformatics knowledge systems. *Journal of Biomedical Informatics*, 41(5):706 – 716, 2008.

[5] The EBI RDF Platform:. http://www.ebi.ac.uk/rdf/.

[6] Michael Ashburner *et al.* Gene Ontology: tool for the unification of biology. *Nature genetics*, 25(1):25–29, 2000.

[7] Adrien Coulet *et al.* Ontology-based knowledge discovery in pharmacogenomics. In *Software Tools and Algorithms for Biological Systems*, pages 357–366. Springer, 2011.

[8] Alison Callahan *et al.* Querying Bio2RDF Linked Open Data with a Global Schema. In *Proceedings of Bio-ontologies SIG*, 2012.

[9] Mark D. Wilkinson *et al.* The Semantic Automated Discovery and Integration (SADI) Web service Design-Pattern, API and Reference Implementation. *J. Biomedical Semantics*, 2:8, 2011.

[10] Pedro Lopes and José Luís Oliveira. COEUS: "semantic web in a box" for biomedical applications. *J. Biomedical Semantics*, 3:11, 2012.

[11] Benjamin M. Good and Mark D. Wilkinson. The Life Sciences Semantic Web is Full of Creeps! *Briefings in Bioinformatics*, 7(3):275–286, 2006.

[12] M. Scott Marshall *et al.* Emerging practices for mapping and linking life sciences data using RDF - A case series. *J. Web Sem.*, 14:2–13, 2012.

[13] Aldo Gangemi *et al.* Automatic typing of DBpedia entities. In *The Semantic Web–ISWC 2012*, pages 65–81. Springer, 2012.

[14] Axel-Cyrille Ngonga Ngomo. Link discovery with guaranteed reduction ratio in affine spaces with minkowski measures. In *The Semantic Web–ISWC 2012*, pages 378–393. Springer, 2012.

[15] Mengling Xu *et al.* Discovering Missing Semantic Relations between Entities in Wikipedia. In *The Semantic Web–ISWC 2013*, pages 673–686. Springer, 2013.

[16] Christian Y. A. Brenninkmeijer *et al.* Computing Identity Co-Reference Across Drug Discovery Datasets. In *Proceedings of SWAT4LS 2013*, 2013.

[17] Bethany Percha *et al.* Discovery and explanation of drug-drug interactions via text mining. In *Pacific Symposium on Biocomputing. Pacific Symposium on Biocomputing*, pages 410–421. World Scientific, 2012.

[18] Jyotishman Pathak *et al.* Mining Anti-coagulant Drug-Drug Interactions from Electronic Health Records Using Linked Data. In *DILS*, pages 128–140. Springer, 2013.

[19] Mathieu d'Aquin *et al.* Combining data mining and ontology engineering to enrich ontologies and linked data. In *Workshop: Knowledge Discovery and Data Mining Meets Linked Open Data-Know@ LOD at ESWC*, volume 2012, 2012.

[20] Luis Antonio Galárraga *et al.* Amie: association rule mining under incomplete evidence in ontological knowledge bases. In *Proceedings of the 22nd international conference on World Wide Web*, pages 413–422. International World Wide Web Conferences Steering Committee, 2013.

[21] Ashwin Srinivasan. The Aleph Manual. Available at http://www.comlab.ox.ac.uk/oucl/research/areas/machlearn/Aleph/, 2007.

[22] Renaud Grisoni *et al.* Méthodologie et outils pour l'extraction de connaissances par Programmation Logique Inductive (PLI) (Poster). In *EGC 2013*, Toulouse, France, 2013.

[23] Clara DM van Karnebeek and Sylvia Stockler. Treatable inborn errors of metabolism causing intellectual disability: a systematic literature review. *Molecular genetics and metabolism*, 105(3):368–381, 2012.

[24] Arno Knobbe *et al.* From Local Patterns to Global Models: The LeGo Approach to Data Mining. In *International Workshop From Local Patterns to Global Models co-located with ECML/PKDD'08*, pages 1–16, Antwerp, Belgium, September 2008.

[25] Mark-A. Krogel *et al.* Comparative evaluation of approaches to propositionalization. In *ILP*, pages 197–214, 2003.