HAL
open science

Data-Interlinking: the Seed of Knowledge Reconciliation in Pharmacogenomics

Pierre Monnin, Amedeo Napoli, Adrien Coulet

To cite this version:

Pierre Monnin, Amedeo Napoli, Adrien Coulet. Data-Interlinking: the Seed of Knowledge Reconciliation in Pharmacogenomics. 2018. hal-01955262

HAL Id: hal-01955262

https://inria.hal.science/hal-01955262v1

Preprint submitted on 14 Dec 2018

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Data-Interlinking: the Seed of Knowledge Reconciliation in Pharmacogenomics

Pierre Monnin[1], Amedeo Napoli[1], and Adrien Coulet[1,2]

[1] Université de Lorraine, CNRS, Inria, LORIA, F-54000 Nancy, France
firstname.lastname@loria.fr
[2] Stanford Center for Biomedical Informatics Research, Stanford University, Stanford, California

Pharmacogenomics (PGx) studies how genomic variations impact drug responses. Knowledge units in PGx have typically the form of ternary relationships *genomic variation – drug – phenotype*, stating that patients having the *genomic variation* and being treated with the *drug* will experience the *phenotype*, which can be the expected outcome of the drug treatment or an adverse effect. For example, one well studied PGx relationship is *G6PD:202A – chloroquine – anemia* which states that patients having the *202A* version of the *G6PD* gene and being treated with *chloroquine* will be more likely to experience *anemia*.

State-of-the-art knowledge in PGx is available in specialized knowledge bases (such as PharmGKB) and in the biomedical literature. PGx relationships may come with a kind of a *truth value*, identifying if they have been thoroughly studied and validated or if they have only been observed on reduced cohorts of patients. On the other hand, PGx knowledge units may be discovered by mining Electronic Health Records (EHRs) of hospitals. Such discovered knowledge units could then be compared with the state of the art in order to confirm poorly validated PGx relationships [1]. Therefore, one major task resides in the *knowledge reconciliation* of these different sources, *i.e.*, identifying when two PGx relationships are in fact referring to the same knowledge unit, if one is a more precise version of the other, if they are related to some extent or if they are different knowledge units. In a way, the knowledge reconciliation task can be seen as an extension of data interlinking.

We arranged the first bricks of such knowledge reconciliation in a preliminary work [2] and extended it recently with a work under revision [3]. The main idea to perform this reconciliation resides in pair-wise comparing the respective sets of drugs, genomic factors and phenotypes involved in relationships. For example, if two PGx relationships involve the same sets of drugs, genomic factors and phenotypes, they represent the same knowledge unit. A PGx relationship can be a more precise version of another for instance if it involves only a subset of drugs, or a more specific phenotype w.r.t. an ontology (*e.g.*, `Heart Block`, a subclass of `Heart Diseases` according to the ontology MeSH). This reasoning on involved sets of components led us to define a set of five *reconciliation rules* [3]. On the left side of a rule, equalities or inclusions between respective sets of components of two PGx relationships are tested and combined with conjunctions and disjunctions. The right side of a rule consists of a link between the two

compared PGx relationships to be added to the knowledge base if and only if the left side is true.

Because of their different origins, PGx knowledge units may be heterogeneously described, in terms of languages, vocabularies and granularities. To address this heterogeneity within our rules, we rely on background knowledge in the form of biomedical ontologies and linked data sets (such as the ones resulting from the Bio2RDF project) where hierarchies of classes and interlinking of individuals are provided. These hierarchies and interlinking allow us to identify when two involved sets of components are formed by the same individuals, or if one set is more precise (or subsumed) by the other set.

For the comparison mechanism to yield as much results as possible, the hierarchies of classes and interlinking provided in the considered background knowledge must be as complete as possible. However, this seems difficult in reality. Indeed, as phenotypes can be combined and adapted, there is no complete ontology or dictionary. For example, complex phenotypes can depend on the taken drug, such as *carbamazepine hypersensitivity*, or express a modified risk, such as *increased risk of Stevens-Johnson syndrome*. This latter phenotype can be considered similar but not strictly identical to *Stevens-Johnson syndrome*. Consequently, relationships involving these phenotypes should be considered as similar but not strictly identical. Such a similarity link, even though not being a strict equivalence, is still of interest in our context of knowledge reconciliation.

Dealing with these similarities and diversities is hard in our symbolic rule-based approach. Even though in [3] we defined one rule to identify similar relationships, it only expresses a static similarity based on the presence of a specific predicate. In this case, numerical methods, by learning some similarity metrics, could help completing the reconciliation results. Such methods could leverage multiple features, *e.g.*, cross-references in linked data sets, chemicals involved in a drug, nested phenotypes, *etc.* Finally, results of numerical methods could somehow enrich the symbolic methods. Indeed, by comparing results from numerical methods with results from symbolic methods, we could identify cases where our rules fail and, therefore, enrich them.

Finally, as aforementioned, knowledge reconciliation both extends and relies on interlinking. That is why the latter can be considered as a seed for the former.

References

1. Coulet, A., Smaïl-Tabbone, M.: Mining electronic health records to validate knowledge in pharmacogenomics. ERCIM News **2016**(104) (2016), http://ercim-news.ercim.eu/en104/special/mining-electronic-health-records-to-validate-knowledge-in-pharmacogenomics
2. Monnin, P., Jonquet, C., Legrand, J., Napoli, A., Coulet, A.: PGxO: A very lite ontology to reconcile pharmacogenomic knowledge units. PeerJ PrePrints **5**, e3140 (2017). https://doi.org/10.7287/peerj.preprints.3140v1
3. Monnin, P., Legrand, J., Husson, G., Ringot, P., Tchechmedjiev, A., Jonquet, C., Napoli, A., Coulet, A.: PGxO and PGxLOD: a reconciliation of pharmacogenomic knowledge of various provenances, enabling further comparison. bioRxiv p. 390971 (2018). https://doi.org/10.1101/390971