Marco Brandizi[1] / Ajit Singh[1] / Christopher Rawlings[1] / Keywan Hassani-Pak[1]

Towards FAIRer Biological Knowledge Networks Using a Hybrid Linked Data and Graph Database Approach

[1] Rothamsted Research, Computational and Analytical Sciences Department, Harpenden, AL5 2JQ, UK, E-mail: marco.brandizi@rothamsted.ac.uk. http://orcid.org/0000-0002-5427-2496, http://orcid.org/0000-0001-6239-9967, http://orcid.org/0000-0003-3702-1633, http://orcid.org/0000-0001-9625-0511.

Abstract:
The speed and accuracy of new scientific discoveries – be it by humans or artificial intelligence – depends on the quality of the underlying data and on the technology to connect, search and share the data efficiently. In recent years, we have seen the rise of graph databases and semi-formal data models such as knowledge graphs to facilitate software approaches to scientific discovery. These approaches extend work based on formalised models, such as the Semantic Web. In this paper, we present our developments to connect, search and share data about genome-scale knowledge networks (GSKN). We have developed a simple application ontology based on OWL/RDF with mappings to standard schemas. We are employing the ontology to power data access services like resolvable URIs, SPARQL endpoints, JSON-LD web APIs and Neo4j-based knowledge graphs. We demonstrate how the proposed ontology and graph databases considerably improve search and access to interoperable and reusable biological knowledge (i.e. the FAIRness data principles).

1 Introduction and Related Work

In 2010, the advent of a data-based society was emphasised by a news magazine with a cover dedicated to "the data deluge" [1]. Collecting, exchanging and processing data and information are ever more fundamental activities to improve our lives, which are impacting fields like businesses [2], [3], [4], manufacturing [5], [6], medicine [7], [8], [9], agriculture [10], [11], [12] and even humanities [13]. This has been made possible by advances in computer science and the spread of Internet standards, which have both provided with the technical means to deal with high amounts of interconnected data and created a culture that favours information exchange and sharing [14], [15]. In science, and particularly in life sciences, these trends have started even before they became popular more widely [16], [17], [18]. In fact, high throughput biotechnologies extract massive amounts of data out of organisms, which need to be turned into useful heterogeneous biological information [19]. Data have become so fundamental in life science and other sciences that a group of stakeholders have published the "FAIR data principles", which establish that data should be Findable, Accessible, Interoperable and Reusable [20], [21]. Regarding the technical aspects, the importance of standards for data modelling and knowledge representation, as well as sharing best practices, have been stressed for years, in life sciences as in other fields [22], [23]. In particular, organisations like the W3C have promoted the idea of leveraging the traditional World Wide Web to create the Semantic Web [24], [25], which makes it possible to publish and consume a web of data, similarly to what we do with the better known web of documents. Linked open data projects have emerged as concrete realisations of such vision [26], where the "open" adjective stresses legal and social aspects and the need to make information freely accessible and reusable. More recently, a number of different formats and technologies, such as JSON-LD [27], [28] or noSQL databases [29], [30] have emerged to model and manage data, which are simpler than highly formal languages such as OWL [31], [32], and based on pre-existing standards and software engineering approaches, such as JSON [33], [34] or REST APIs [35]. Knowledge graphs are one prominent example of such recent trends, which, significantly, are being promoted by commercial organisations, after having started as tools to support their own business [36], [37]. According to [38], a

knowledge graph is a large heterogeneous knowledge base, modelled through graphs and ontologies, which derives new knowledge from existing data sets such as Freebase [39] or Wikidata [40], by means of various inference techniques, including automatic reasoning, natural language processing, statistical analysis and user crowdsourcing [41], [42]. In addition to this definition, here we suggest that these knowledge bases tend to be based on ontologies and schemas that are more informal than what one can see in a traditional expert system, or in a typical description logic-based ontology (and hence a typical OWL-based data set). On the one hand, well formalised data allow for precise information retrieval and extraction, as well as advanced forms of automated logical reasoning [43]. However, dealing with very formal data models is difficult, both for software developers and end users, including when they are domain experts [44]. Such difficulty is increased in the context of the web, where large amounts of data imply large amounts of data noise and syntax and format differences [45].

1.1 The KnetMiner Use Case

In the last decade, our group has been strongly involved in the development of software for data integration and biological knowledge discovery. The KnetMiner software suite is primarily based on web components and applications to mine large, heterogeneous knowledge networks called genome-scale knowledge networks (GSKN [46], [47]). A user can start with search keywords and then explore genes that are significantly related to the search inputs, according to networks of connected knowledge, including encoded proteins, biological pathways, scientific literature, diseases and phenotypes. KnetMiner is presently based on the Ondex software platform [48], a data integration framework that allows for importing multiple data sources into a unified data model, viewing, analysing and transforming imported data in a desktop application and exporting data integrated this way to an XML file, which is based on the OXL format, defined by means of XML Schema [49]. While data management through Ondex and the OXL format is relatively easy, Semantic Web technologies, linked data principles and graph databases would give interesting further opportunities. In order to take advantage of them, in this paper we present BioKNO, a new OWL-based application ontology that leverages our experience with Ondex, OXL and KnetMiner to define a data model about what we call biological knowledge networks. A GSKN can be considered a particular case of biological network, i.e. knowledge graphs which, while being focused on the life science domain, aim at being of general use for this and other domains, as well as modelling data with a good balance between formalisation and practical usefulness, especially in a life science community that frequently exchanges data via the web and intensely relies on unstructured or semi-structured data.

2 Architecture and Implementation

2.1 The Bio-Knowledge Network Ontology (BioKNO)

In order to address data integration and access needs, we have developed the Bio-Knowledge Network Ontology (BioKNO, pronounced "bio-know" [50]). This is a lightweight and general ontology, which allows for modelling a wide variety of life science entities in a simple way. Given its scope, BioKNO can be considered an application ontology, i.e. an ontology more focused on supporting the specific needs of a community and a set of software applications, rather than a general purpose foundational/upper ontology [51].

As shown in Figure 1, BioKNO allows for the classification of entities of interest into the general top-level class of *Concept* and the identification of relations of interest that connect concepts by means of the similarly general container *relatedConcept*. It is expected that specific kinds of concepts are defined by subclassing *Concept* and, similarly, specific relations extend *relatedConcept*. As it is common practice, new relations, which are OWL object properties [52], can define the domain and range of linked concepts.
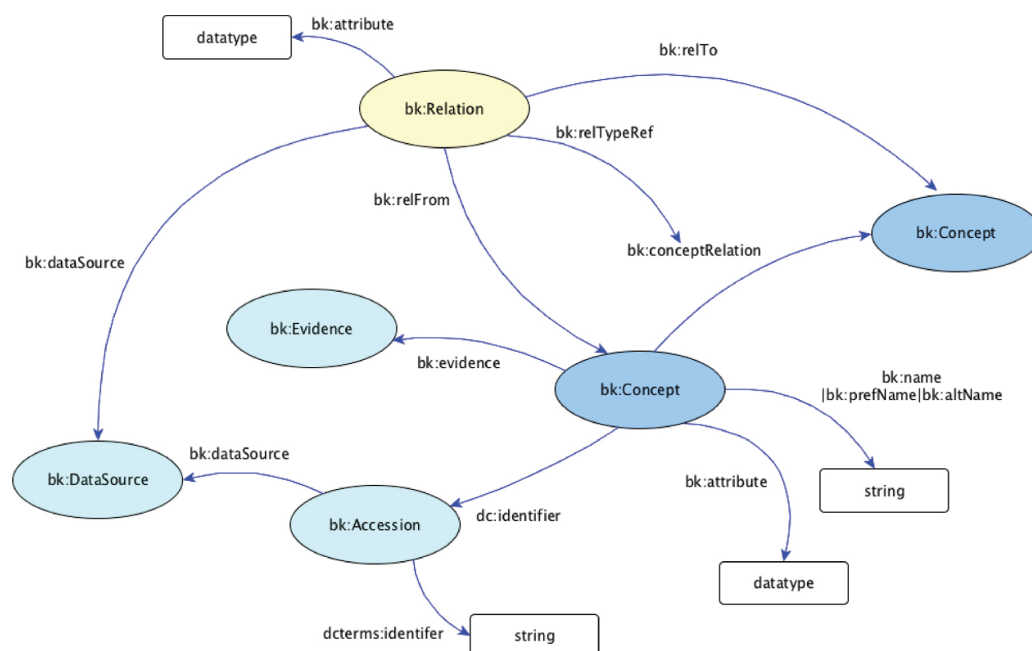
Figure 1: The top-level organisation of the BioKNO ontology.
Specific biological entities are defined as subclasses of *bk:Concept* and relation types (i.e. OWL object properties) between them are based on subproperties of *bk:conceptRelation*. *bk:Relation* can be used to model reified relations. Both concepts and reified relations can have *bk:attribute* and other elements attached.

Concepts can be characterised by subproperties of the attribute property, which is an OWL datatype property [52], i.e. it and its subproperties have plain types such as numbers or strings as range. This way, concepts can be characterised by plain values like the abstract of a publication, or the genome coordinates of a gene. Reified relations [53] are another general construct that we introduce in order to support binary relations between concepts that have additional plain attributes (e.g. a numerical score, a textual annotation), or additional links (e.g. the public database of provenance). The same attribute properties that we have defined for concepts can be used for reified relations, thus leveraging the same features, such as specification of domains and ranges or hierarchical information (e.g. a PMID attribute is a particular type of identifier). While this basic and lightweight core is very generic and, in principle, it might be used to model any type of knowledge, we have extended it with a number of classes and properties about entities that are common in biological knowledge bases [54], including molecular entities such as genes or proteins, relations like *expressed_by* or *catalyzes*, and attributes such as *sequence* or *p*-value. For all common concept types, relation types and attribute names that we have defined, we have specified basic ontological properties, such as hierarchical relations (e.g. *published_in* is a subproperty of *related_to*), or domains and ranges (e.g. *PMID*, *abstract*, *authors* have *Publication* as domain, *p*-value has *xsd:double* as range).

2.2 BioKNO by Examples

In order to show how to use BioKNO to model network knowledge, in the supplementary document 1 we have outlined examples taken from a simplified (yet significant) selection of real data. Example 1 shows a biological pathway taken from WikiPathways and one of its participant protein modelled with BioKNO. WikiPathways is a database of biological pathways maintained by the scientific community. The WikiPathways RDF in BioPAX format [55] is provided as part of the monthly releases and contains curated pathway information, as well as imports from sources like Reactome [56]. This and other examples in the document illustrate the easy integration of pathway information and related knowledge (e.g. ontologies, literature), made possible by BioKNO. For instance, complex chains of relations that link proteins to pathways in BioPAX are summarised by direct *participates_in* relations, which is sufficient for many knowledge discovery applications (e.g. find pathways relevant to a gene list). This conversion can be made using the SPARQL query language, by simply defining BioPAX reactions as graph patterns and translating them into another graph pattern, in a CONSTRUCT query [57]. In example 2, we show how to use BioKNO to view a complex and formal ontology like Gene Ontology as a concept scheme [58], another type of simplified data model, which is useful in many applications. Finally, in example 3 we describe the use of attributes in BioKNO, including their use as relation attributes, by means of reified relations.
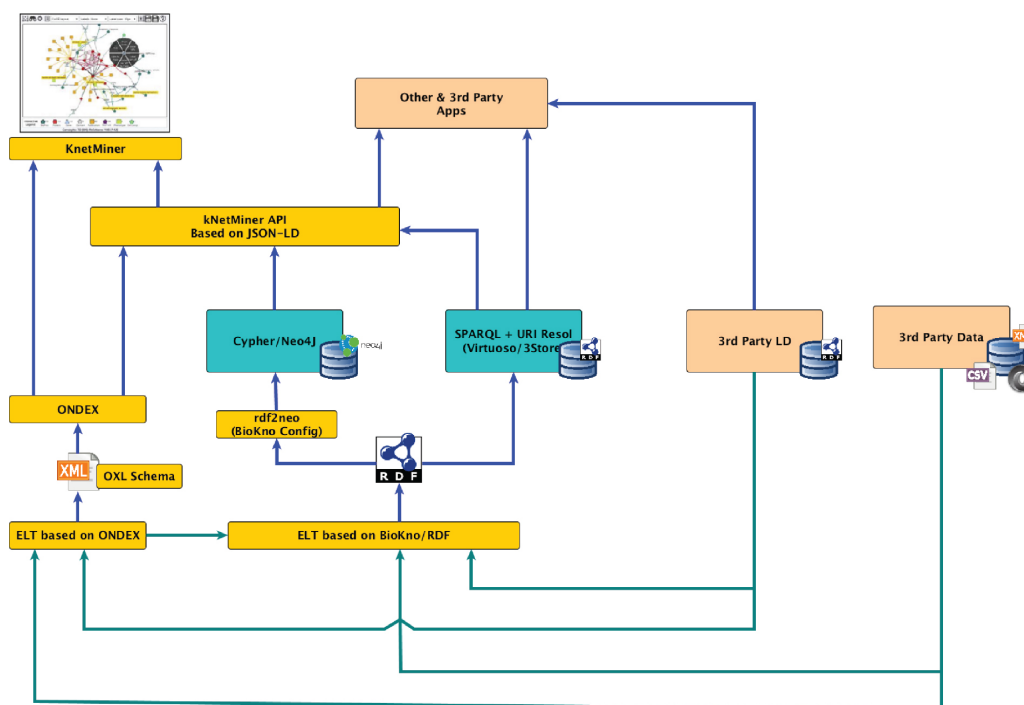
Figure 2: The new architecture designed for the KnetMiner ecosystem.
BioKNO-based modelling powers both data acquisition (extraction, loading, transformation, or ELT) and data querying from our and 3rd-party applications. RDF serves open data publishing and integration with other data sets.

2.3   New KnetMiner Architecture and Applications

Figure 2 shows how we are re-designing the architecture for the aforementioned KnetMiner application and ecosystem. BioKNO, and hence OWL and RDF, will be at the centre of it, progressively replacing the Ondex XML format. The new format will be both a reference for converting external data of interest into a unified data format, and a source of queries, data access and exchange for applications. We have achieved a first concrete result by reviewing past work [59], and developing an Ondex plug-in to export OXL data as RDF that instantiates our ontology [60].

3   Applications

In this section we describe applications where BioKNO is either being used, or planned to be used, or potentially useful. As a confirmation of the flexibility of a graph-based, linked data-based approach, these applications are either based on the KnetMiner architecture outlined above, or independent on it and relying on third party tools and standard formalisms. Moreover, we mention in which ways our applications can address the FAIR requirements described by the FORCE11 paper [20].

## 3.1   BioKNO as Exchange Format for Graph Databases

A natural application of BioKNO consists of using the RDF obtained from Ondex to populate an RDF triple store based on Virtuoso [61]. As it is common in the linked data context, we plan to make this public, in the form of both an endpoint to query data by means of the Semantic Web language SPARQL [62] and via URI resolution [63]. The latter allows for the *Findability* in FAIR. In particular, we are going to publish proper dataset metadata, formalised through the VoID [64] and Dublin Core [65] vocabularies. Since our data are essentially fully open, adopting these vocabularies will also make it easy to publish machine-actionable information related to data licensing, as recommended by the FAIR principles. Adopting the SPARQL and RDF/OWL standards makes data *Accessible* and *Interoperable*. In addition to triple store use, we have setup a Neo4j server to make our knowledge graphs accessible through the Cypher query language [66]. Importantly, the Neo4j graph database is based on the same RDF that is exported from Ondex, in order to keep both databases aligned to the same data

model. This improves data FAIR *Reusability* by means of diversification toward multiple formats of a unified data model. As shown in Figure 1, this diversification includes exposing data access through web service APIs and the JSON-LD language. The conversion from RDF to the Neo4j-based linked property graph model is made possible through rdf2neo [67], a tool that we have developed to convert RDF data into suitable Cypher instructions. In contrast to [68], our mapping is highly configurable by means of SPARQL queries, which associate the RDF input to Cypher entities (e.g. *rdf:type* statements to node labels, triples to Cypher relations), making rdf2neo highly flexible. Cypher is a declarative, SQL-inspired language for describing patterns in graphs visually using an ascii-art syntax. It is at the base of Neo4j and is becoming increasingly popular in similar graph databases [69]. In our preliminary tests, we found Cypher as particularly easy, compact and fast to query large genome-scale knowledge graphs for gene-phenotype related information.

3.2 SPARQL as Data Transformation and Integration Language

To produce the BioKNO RDF from WikiPathways RDF, we use RDF transformations based on SPARQL CONSTRUCT queries [70] and CSV conversions based on the TARQL tool, which, again, exploits SPARQL to map table structures to RDF [71]. These transformations are generally easy to write and require only basic knowledge of SPARQL. This new approach of transforming data into graphs compares favourably from an efficiency and ease-of-use standpoint to the earlier Ondex approach, which required developing a new Java-based Ondex plugin whenever a new data format was to be introduced into the Ondex framework. Using a standard like RDF to ease data transformation operations contributes to the *Reuse* in the FAIR principles.

```
// Blue path, () for 'any node', -()- (without '>') for 'any direction', [r1|r2] to OR-match two relations
MATCH  (g:Gene)-->()-[:ortho]->()<-[:enc]-()-[:physical|:genetic]-(g1:Gene)
WHERE 'RIN4' IN ( g.prefName + g.altName )
RETURN
  g.identifier, 'Gene' AS evidenceType, g1.prefName + g1.altName AS evidenceTitles, null as evidenceDescr,
  2 AS score // gene evidence entities receives a first score weight based on their types or topology
// Green path, [*1..2] for any path of 1-2 length, composed with whatever relation
UNION MATCH (g:Gene) - [*1..2] -> (pub:Publication)
WHERE 'RIN4' IN ( g.prefName + g.altName )
RETURN
  g.identifier, 'Publication' AS evidenceType, pub.AbstractHeader AS evidenceTitles,
  pub.Abstract AS evidenceDescr, 1 AS score
```
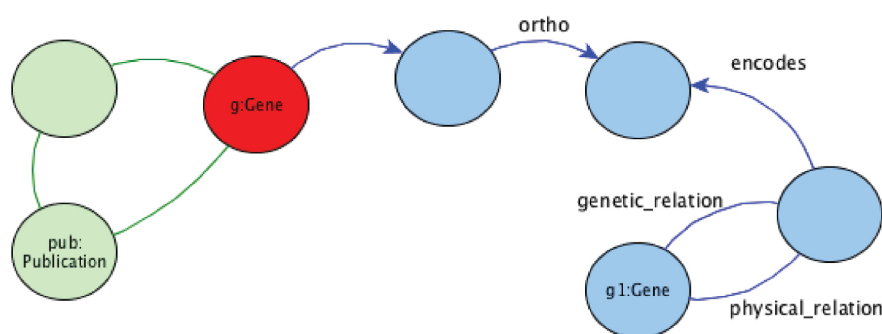

Figure 3: Using Cypher to query genome-scale knowledge networks. The query corresponds to the graph pattern on the bottom.

3.3 Cypher as a Search Language for Biological Connections

As an example for the utility of Cypher, consider the KnetMiner approach to search and rank genes based on gene evidence networks [46]. The main idea of the approach consists of finding genes based on associations with entities (e.g. phenotypes) and documents (e.g. scientific articles) in which search keywords occur in a statistically significant way. KnetMiner makes extensive use of graph queries in the form of what we call semantic motif searches, i.e. graph pattern-based searches employing the declaration of well-defined entity types and their linking relations (e.g. *Gene -> expresses -> Protein -> published_in -> Publication*). By having generated a Neo4j-based version of our knowledge networks, we are now in the position to redesign the motif-based component of KnetMiner to support Cypher as the language for specifying gene-evidence graph patterns. As a pattern search language, Cypher is significantly more expressive than our existing motif search language and, as another advantage, it is getting a de-facto standard for graph databases. Figure 3 shows an example cypher query that consists of a UNION statement and constraints about gene names and maximum path length. Fi-

nally, endpoints like Neo4j make it possible to easily access our knowledge networks from different languages such as Python and R [72], two popular languages among data scientists.

3.4   Talking to the Rest of the World

Following Semantic Web and linked data principles, we have started to leverage BioKNO to integrate our biological knowledge networks with existing relevant data. Commonly used entities were aligned to relevant entities in ontologies/schemas like BioPAX [55], schema.org [73], Semantic Science Integrated Ontology (SIO [74]), Relation Ontology (RO [75]), Basic Formal Ontology (BFO [76]), SKOS [58] and Dublin Core [65]. This mapping was made manually, relying on bioinformatics expertise. Results are available in our ontology repository [77]. The supplementary document 2 summarises the most relevant entities that were mapped this way. The document also reports an example showing how such mapping can be useful, which consists of a SPARQL query that combines WikiPathways data and pathways in Pathway Commons [78] in which the same proteins participate. This example shows that data from different sources can be accessed by means of common data models and ontologies, even when more specific models are used locally. For example, the mentioned query uses entities from the BioPAX vocabulary, which are matched to our data by means of OWL-based automated reasoning applied to our ontological mappings. Furthermore, this highlights that, as in similar cases, aligning BioKNO to existing common ontologies is a major contribution to the FAIR principle of *Interoperability* and *Findability*.

4   Discussion and Conclusions

Proper data models and data standards are key to realise the goals and hopes of the FAIR principles, which, in turn, are very important for improving the digital access to scientific knowledge, biological knowledge in particular, and the collaboration to produce and reuse such knowledge. In recent years, solutions like knowledge graphs and technologies like graph databases have become popular in production-level and commercial contexts. In part, traditionally academic approaches like linked data have moved out of the academia and have enriched these new developments, with examples like JSON-LD or graph-based frameworks wrapping RDF triple stores [79]. Our work shows how using "traditional" Semantic Web languages like OWL to build a biological knowledge network can help in building an ecosystem of applications that leverage common life science information. In doing so, balancing a good degree of formalisation with more informal approaches eases the development of a FAIR-based platform like KnetMiner with data imports, information retrieval, data republishing for re-use and integration of/into external data sets, so that we can promote collaborative open data science. While the use of RDF-based data publishing contributes to the realisation of findable, accessible and reusable data (and ontology alignment contributes mostly to interoperability), BioKNO has been designed with the aim of being a reference model beyond the initial choice of the OWL language, which, FAIR-wise, makes them even more interoperable and reusable. In fact, we plan to translate the same data model into other formats, such JSON-LD, similarly to what has been done in related projects [80]. Using our RDF/OWL model to create a corresponding translation for the Neo4j database is a further example of such flexibility with multiple data formats, which compares to projects that uses Neo4j to serve biological network data [81], [82], [83], where data access is more tightly coupled to the storage backend. In our experience so far, there are significant practical differences between triple stores coupled with the SPARQL query language and more recent graph databases like Neo4j, based on labelled property graphs [84] driven by a language like Cypher. This type of graph-based databases have complementary, rather than alternative characteristics to Semantic Web-based systems, which lead us to decide to support both types of endpoints. We base these considerations on a first assessment of these pros and cons [85] and we also plan to undertake more thorough evaluations in future. In the supplementary document 3 we report a summary of preliminary results that outlines features like compactness of the Cypher language, attribute-supporting relations in Neo4j (which avoids the need for reified relations), higher expressivity of triple stores, native support to data format standardisation in the Semantic Web world.

Similarly to already mentioned knowledge graph projects, our lightweight modelling approach has its own limits, in particular regarding the trade-off between precise formal representations of reality and flexibility in modelling biological knowledge. To clarify this, consider the Gene Ontology case described above (example 2 in Supplementary Document 1): the simplification consisting in the use of straight *part_of* statements between OWL classes (e.g. *obo:GO_0030015* "CCR4-NOT core complex" *part_of GO_0030014* "CCR4-NOT complex") is not formally precise/correct in OWL, since it normally requires to describe relations between instances of classes (i.e. *obo:GO_0030015 rdfs:subClassOf [bk:part_of some obo:GO_0030014]*). Our simpler representation is still correct in OWL-2, thanks to the punning mechanism [86], which can be seen as a way to enrich two URIs with the

interpretation that they refer not only to ontological classes, but also to exemplary protein complexes, which are instances of the respective classes. However, while still correct, such simplification limits the OWL reasoning capabilities, in the sense that, in absence of more proper OWL definitions, no OWL reasoner will infer that triple sets like: *{prot1 bk:part_of prot2; prot2 rdf:type obo:GO_0030014}* lead to the conclusion that *prot1* is an instance of *GO_0030015*. Limits like this can be overcome with mechanisms such as inference rules that translate from informal models like BioKNO to OWL (and vice-versa). Furthermore, we are reviewing our defined classes and properties following a methodology developed in previous work to address these and other issues [59]. Another compromise that we are aware of is about defining our own vocabulary, even for those entities that are well known in other schemas and ontologies (e.g. *bk:name* vs. *schema:name*, *bk:part_of* vs. *obo:BFO_0000051*, *bk:Path* vs. *biopax:Pathway*). Usually adopting terms from the already existing vocabularies is cleaner and preferable in the context of linked open data, but not for a community of end-users and developers who are not specialists of knowledge representation or Semantic Web standards, especially if instead they are already familiar with a given, more scope-narrowed terminology. Indeed, they would not like to have to deal with many different schemas and vocabularies at the same time, a task that can be difficult even for experts [87]. Mapping users' preferred terms to existing standards is a good way to balance their needs with the more general goal of making data more widely available and re-usable (as we say in this article title, *FAIRer*). Finally, we have noticed some performance limitations in running input/output operations when working with RDF data compared to the original OXL format: data load and export times are smaller when ad-hoc Ondex components are used for this (up to a 2x factor, when considering the time needed to populate Virtuoso or Neo4j). This is mostly because Ondex has been optimised for the OXL format, while general RDF software libraries are used to process RDF data files (which are not tailored to BioKNO). Since these times are still in the order of few minutes, even for the largest datasets (million of nodes/relations), we see this limit as a reasonable price to pay in exchange for more flexibility with data processing and data accessibility. To improve the performance with RDF, we might consider optimised formats like HDT [88] in future.

As further future work, we plan to support more RDF/OWL standards, as well as contributing to their development. In particular, we are interested in the ongoing work with bioschemas [89], which has goals similar to ours about defining a simple schema about biological entities. Furthermore, we plan to realise data and application integration based on JSON-LD and web APIs, for example, the BrAPI [90]. This will also be a base to promote the development of third-party applications, including those using translations of our OWL/RDF data to JSON-LD and Neo4j.

Acknowledgements

Funding

Conflict of interest statement: Authors state no conflict of interest. All authors have read the journal's Publication ethics and publication malpractice statement available at the journal's website and hereby confirm that they comply with all its parts applicable to the present scientific work.

References

[1] The data deluge [Internet]. The Economist; 2010. Available from: https://www.economist.com/node/15579717.
[2] Bennett M. The financial industry business ontology: best practice for big data. J Bank Regul. 2013;14:255–68.
[3] O'Riain S, Curry E, Harth A. XBRL and open data for global financial ecosystems: a linked data approach. Int J Account Inf Syst. 2012;13:141–62.

[4] Third A, Domingue J. Linked Data Indexing of Distributed Ledgers. Proc 26th Int Conf World Wide Web Companion [Internet]. Republic and Canton of Geneva, Switzerland: International World Wide Web Conferences Steering Committee; 2017 [cited 2018 Mar 7]. p. 1431–6. Available from: https://doi.org/10.1145/3041021.3053895.

[5] Laursen K, Salter A. Open for innovation: the role of openness in explaining innovation performance among U.K. manufacturing firms. Strateg Manag J. 2006;27:131–50.

[6] Lee J, Lapira E, Bagheri B, Kao H. Recent advances and trends in predictive manufacturing systems in big data environment. Manuf Lett. 2013;1:38–41.

[7] Hassanzadeh O, Kementsietsidis A, Lim L, Miller RJ, Wang M. LinkedCT: A Linked Data Space for Clinical Trials. ArXiv09080567 Cs [Internet]. 2009 [cited 2018 Mar 7]; Available from: http://arxiv.org/abs/0908.0567.

[8] Samwald M, Jentzsch A, Bouton C, Kallesøe CS, Willighagen E, Hajagos J, et al. Linked open drug data for pharmaceutical research and development. J Cheminformatics. 2011;3:19.

[9] Chen Y, Argentinis JE, Weber G. IBM Watson: how cognitive computing can be applied to big data challenges in life sciences research. Clin Ther. 2016;38:688–701.

[10] Akhtar P, Tse YK, Khan Z, Rao-Nicholson R. Data-driven and adaptive leadership contributing to sustainability: global agri-food supply chains connected with emerging markets. Int J Prod Econ. 2016;181:392–401.

[11] Venkatesan A, El Hassouni N, Phillipe F, Pommier C, Quesneville H, Ruiz M, et al. Exposing French agronomic resources as Linked Open Data. Ing Connaiss IC2016 – Workshop Ovive [Internet]. Montpellier, France; 2016 [cited 2018 Mar 7]. Available from: https://hal.archives-ouvertes.fr/hal-01411759.

[12] Caracciolo C, Stellato A, Morshed A, Johannsen G, Rajbhandari S, Jaques Y, et al. The AGROVOC Linked Dataset. Semantic Web. 2013;4:341–8.

[13] Barbera M. Linked (open) data at web scale: research, social and engineering challenges in the digital humanities. JLIS It. 2013;4:91.

[14] Pohorec S, Zorman M, Kokol P. Analysis of approaches to structured data on the web. Comput Stand Interfaces. 2013;36:256–62.

[15] Allen M. What was Web 2.0? Versions as the dominant mode of internet history. New Media Soc. 2013;15:260–75.

[16] Wang X, Gorlitsky R, Almeida JS. From XML to RDF: how semantic web technologies will change the design of 'omic' standards. Nat Biotechnol. 2005;23:1099–103.

[17] Ruttenberg A, Clark T, Bug W, Samwald M, Bodenreider O, Chen H, et al. Advancing translational research with the Semantic Web. BMC Bioinformatics. 2007;8:S2.

[18] Belleau F, Nolin M-A, Tourigny N, Rigault P, Morissette J. Bio2RDF: towards a mashup to build bioinformatics knowledge systems. J Biomed Inform. 2008;41:706–16.

[19] Lesk A. Introduction to bioinformatics. Oxford, UK: OUP, 2013.

[20] Wilkinson MD, Dumontier M, Aalbersberg IjJ, Appleton G, Axton M, Baak A, et al. The FAIR Guiding Principles for scientific data management and stewardship. Sci Data. 2016;3:160018.

[21] Mons B, Neylon C, Velterop J, Dumontier M, da Silva Santos LOB, Wilkinson MD. Cloudy, increasingly FAIR; revisiting the FAIR Data guiding principles for the European Open Science Cloud. Inf Serv Use. 2017;37:49–56.

[22] Smith B, Ashburner M, Rosse C, Bard J, Bug W, Ceusters W, et al. The OBO Foundry: coordinated evolution of ontologies to support biomedical data integration. Nat Biotechnol. 2007;25:1251–5.

[23] Bovalis K, Peristeras V, Abecasis M, Abril-Jimenez RM, Rodríguez MA, Gattegno C, et al. Promoting Interoperability in Europe's E-Government. Computer. 2014;47:25–33.

[24] Berners-Lee T, Hendler J, Lassila O. The semantic web. Sci Am. 2001;284:34–43.

[25] Antoniou G, Van Harmelen F. A semantic web primer. Cambridge, UK: MIT press, 2004.

[26] Bizer C, Heath T, Berners-Lee T. Linked data-the story so far. Int J Semantic Web Inf Syst. 2009;5:1–22.

[27] Lanthaler M, Gütl C. On using JSON-LD to create evolvable RESTful services. Proc Third Int Workshop RESTful Des. ACM, 2012. p. 25–32.

[28] Lanthaler M. Creating 3rd generation web APIs with hydra. Proc 22nd Int Conf World Wide Web. ACM, 2013. p. 35–8.

[29] Cattell R. Scalable SQL and NoSQL data stores. Acm Sigmod Rec. 2011;39:12–27.

[30] Han J, Haihong E, Le G, Du J. Survey on NoSQL database. Pervasive Comput Appl ICPCA 2011 6th Int Conf On. IEEE; 2011. p. 363–6.

[31] McGuinness DL, Van Harmelen F. OWL web ontology language overview. W3C Recomm. 2004;10:2004.

[32] Motik B, Grau BC, Horrocks I, Wu Z, Fokoue A, Lutz C, et al. OWL 2 web ontology language profiles. W3C Recomm. 2009;27:61.

[33] Crockford D. Introducing json. Available https://www.json.org, 2009.

[34] Bray T. The javascript object notation (json) data interchange format. 2017.

[35] Dragoni N, Giallorenzo S, Lafuente AL, Mazzara M, Montesi F, Mustafin R, et al. Microservices: yesterday, today, and tomorrow. Present Ulterior Softw Eng. Springer, 2017. p. 195–216.

[36] Dong X, Gabrilovich E, Heitz G, Horn W, Lao N, Murphy K, et al. Knowledge vault: a web-scale approach to probabilistic knowledge fusion. ACM Press, 2014 [cited 2018 Feb 22]. p. 601–10. Available from: http://dl.acm.org/citation.cfm?doid=2623330.2623623.

[37] Jesse W, Paul T. Facebook Linked Data via the Graph API. Semantic Web. 2013;4:245–50.

[38] Ehrlinger L, Wöß W. Towards a definition of knowledge graphs. Semant Posters Demos SuCCESS. 2016.

[39] Bollacker K, Evans C, Paritosh P, Sturge T, Taylor J. Freebase: a collaboratively created graph database for structuring human knowledge. Proc 2008 ACM SIGMOD Int Conf Manag Data. AcM, 2008. p. 1247–50.

[40] Vrandečić D, Krötzsch M. Wikidata: a free collaborative knowledgebase. Commun ACM. 2014;57:78–85.

[41] Gabrilovich E, Usunier N. Constructing and Mining Web-scale Knowledge Graphs. ACM Press; 2016 [cited 2018 Feb 22]. p. 1195–7. Available from: http://dl.acm.org/citation.cfm?doid=2911451.2914807.

[42] Rospocher M, van Erp M, Vossen P, Fokkens A, Aldabe I, Rigau G, et al. Building event-centric knowledge graphs from news. Web Semant Sci Serv Agents World Wide Web. 2016;37–38:132–51.

[43] Baader F, Horrocks I, Sattler U. Chapter 3 Description Logics. In: van Harmelen F, Lifschitz V, Porter B, editors. Found Artif Intell [Internet]. Elsevier, 2008 [cited 2018 Mar 7]. p. 135–79. Available from: http://www.sciencedirect.com/science/article/pii/S1574652607030039.

[44] Malone icbo2017 keynote [Internet]. Available from: https://www.slideshare.net/JamesMalone5/malone-icbo2017-keynote.

[45] Cai L, Zhu Y. The challenges of data quality and data quality assessment in the big data era. Data Sci J [Internet]. 2015 [cited 2018 Mar 7];14. Available from: http://datascience.codata.org/articles/10.5334/dsj-2015-002/.

[46] Hassani-Pak K. KnetMiner – An integrated data platform for gene mining and biological knowledge discovery [PhD Thesis]. Universität Bielefeld, 2017.

[47] Hassani-Pak K, Rawlings C. Knowledge discovery in biological databases for revealing candidate genes linked to complex phenotypes. J Integr Bioinforma. 2017;14.

[48] Köhler J, Baumbach J, Taubert J, Specht M, Skusa A, Rüegg A, et al. Graph-based analysis and visualization of experimental results with ONDEX. Bioinformatics. 2006;22:1383–90.

[49] Taubert J, Sieren KP, Hindle M, Hoekman B, Winnenburg R, Philippi S, et al. The OXL format for the exchange of integrated datasets. J Integr Bioinforma. 2007;4:27–40.

[50] BioKNO, The Biological Knowledge Network Ontology [Internet]. Available from: https://github.com/Rothamsted/bioknet-onto.

[51] Menzel C. Reference Ontologies – Application Ontologies: Either/or or Both/And?.

[52] OWL 2 Web Ontology Language Primer (Second Edition) [Internet]. Available from: https://www.w3.org/TR/owl2-primer/.

[53] Defining N-ary Relations on the Semantic Web [Internet]. Available from: https://www.w3.org/TR/swbp-n-aryRelations/.

[54] BioKNO extension to define common biological entities [Internet]. Rothamsted Bioinformatics; 2018 [cited 2018 Mar 14]. Available from: https://github.com/Rothamsted/bioknet-onto/blob/master/bk_ondex.owl.

[55] Demir E, Cary MP, Paley S, Fukuda K, Lemer C, Vastrik I, et al. The BioPAX community standard for pathway data sharing. Nat Biotechnol. 2010;28:935.

[56] Fabregat A, Jupe S, Matthews L, Sidiropoulos K, Gillespie M, Garapati P, et al. The Reactome Pathway Knowledgebase. Nucleic Acids Res. 2018;46:D649–55.

[57] bioknet-onto: modelling of BMP/Human pathway [Internet]. Rothamsted Bioinformatics; 2018 [cited 2018 Mar 12]. Available from: https://github.com/Rothamsted/bioknet-onto/tree/master/examples/bmp_reg_human.

[58] Miles A, Matthews B, Wilson M, Brickley D. SKOS core: simple knowledge organisation for the web. In International Conference on Dublin Core and Metadata Applications, Sep 12; 2005. pp. 3–10.

[59] Splendiani A, Rawlings CJ, Kuo S-C, Stevens R, Lord P. Lost in Translation: Data Integration Tools Meet the Semantic Web (Experiences from the Ondex Project). Recent Prog Data Eng Internet Technol [Internet]. Springer, Berlin, Heidelberg; 2012 [cited 2018 Mar 7]. p. 87–97. Available from: https://link.springer.com/chapter/10.1007/978-3-642-28798-5_13.

[60] ONDEX rdf-export-2 plug-in [Internet]. Rothamsted Bioinformatics; 2017 [cited 2018 Mar 12]. Available from: https://github.com/Rothamsted/ondex-knet-builder/tree/master/modules/rdf-export-2.

[61] Erling O, Mikhailov I. RDF Support in the Virtuoso DBMS. In: Pellegrini T, Auer S, Tochtermann K, Schaffert S, editors. Networked Knowl – Networked Media [Internet]. Berlin, Heidelberg: Springer Berlin Heidelberg; 2009 [cited 2018 Mar 6]. p. 7–24. Available from: http://link.springer.com/10.1007/978-3-642-02184-8_2.

[62] Apache Jena – SPARQL Tutorial [Internet]. [cited 2018 Mar 6]. Available from: https://jena.apache.org/tutorials/sparql.html.

[63] Heath T, Bizer C. Principles of Linked Data. Synth Lect Semantic Web Theory Technol [Internet]. 2011 [cited 2018 Mar 6]. p. 1–136. Available from: http://www.morganclaypool.com/doi/abs/10.2200/S00334ED1V01Y201102WBE001.

[64] Alexander K, Hausenblas M. Describing linked datasets – on the design and usage of void, the 'vocabulary of interlinked datasets. Linked Data Web Workshop LDOW 09 Conjunction 18th Int World Wide Web Conf WWW 09. 2009.

[65] Weibel S. The dublin core: a simple content description model for electronic resources. Bull Am Soc Inf Sci Technol. 1997;24:9–11.

[66] Vukotic A. Neo4j in action. Shelter Island, NY: Manning Publications Co; 2015.

[67] rdf2neo: tools to convert/load RDF into Neo4j [Internet]. Rothamsted Bioinformatics; 2018 [cited 2018 Mar 6]. Available from: https://github.com/Rothamsted/rdf2neo.

[68] Barrasa J. Importing RDF data into Neo4j [Internet]. 2016 [cited 2018 Mar 9]. Available from: https://jesusbarrasa.wordpress.com/2016/06/07/importing-rdf-data-into-neo4j/.

[69] Marton J, Szárnyas G, Varró D. Formalising openCypher Graph Queries in Relational Algebra. Adv Databases Inf Syst [Internet]. Springer, Cham; 2017 [cited 2018 Mar 7]. p. 182–96. Available from: https://link.springer.com/chapter/10.1007/978-3-319-66917-5_13.

[70] Appreciating SPARQL CONSTRUCT more – bobdc.blog [Internet]. [cited 2018 Mar 6]. Available from: http://www.snee.com/bobdc.blog/2009/09/appreciating-sparql-construct.html.

[71] tarql: SPARQL for Tables: Turn CSV into RDF using SPARQL syntax [Internet]. Tarql; 2018 [cited 2018 Mar 6]. Available from: https://github.com/tarql/tarql.

[72] Neo4j from R [Internet]. Neo4j Graph Database Platf. [cited 2018 Mar 6]. Available from: https://neo4j.com/developer/r/.

[73] Guha RV, Brickley D, Macbeth S. Schema.Org: evolution of structured data on the web. Commun ACM. 2016;59:44–51.

[74] Dumontier M, Baker CJ, Baran J, Callahan A, Chepelev L, Cruz-Toledo J, et al. The Semanticscience Integrated Ontology (SIO) for biomedical research and knowledge discovery. J Biomed Semant. 2014;5:14.

[75] Smith B, Ceusters W, Klagges B, Köhler J, Kumar A, Lomax J, et al. Relations in biomedical ontologies. Genome Biol. 2005;6:R46.

[76] Arp R, Smith B, Spear AD. Building ontologies with basic formal ontology. MA, USA: MIT Press, 2015.

[77] bk_mappings.ttl [Internet]. Rothamsted Bioinformatics; 2018 [cited 2018 Mar 6]. Available from: https://github.com/Rothamsted/bioknet-onto/blob/master/bk_mappings.ttl.

[78] Cerami EG, Gross BE, Demir E, Rodchenkov I, Babur Ö, Anwar N, et al. Pathway Commons, a web resource for biological pathway data. Nucleic Acids Res. 2011;39:D685–90.

[79] Apache TinkerPop [Internet]. [cited 2018 Mar 6]. Available from: http://tinkerpop.apache.org/providers.html.

[80] González-Beltrán A, Maguire E, Sansone S-A, Rocca-Serra P. linkedISA: semantic representation of ISA-Tab experimental metadata. BMC Bioinformatics. 2014;15:S4.

[81] Fabregat A, Korninger F, Viteri G, Sidiropoulos K, Marin-Garcia P, Ping P, et al. Reactome graph database: efficient access to complex pathway data. PLoS Comput Biol [Internet]. 2018 [cited 2018 Mar 15];14. Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5805351/.

[82] Summer G, Kelder T, Ono K, Radonjic M, Heymans S, Demchak B. cyNeo4j: connecting Neo4j and Cytoscape. Bioinformatics. 2015;31:3868–9.

[83] Lysenko A, Roznovăţ IA, Saqi M, Mazein A, Rawlings CJ, Auffray C. Representing and querying disease networks using graph databases. BioData Min [Internet]. 2016 [cited 2018 Mar 15];9. Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4960687/.

[84] RDF Triple Stores vs. Labeled Property Graphs: What's the Difference? [Internet]. Neo4j Graph Database Platf. 2017 [cited 2018 Mar 6]. Available from: https://neo4j.com/blog/rdf-triple-store-vs-labeled-property-graph-difference/.

[85] Brandizi M. ONDEX & GrapH DBs [Internet]. 2017 [cited 2018 Mar 6]. Available from: https://github.com/marco-brandizi/odx_neo4j_converter_test.

[86] Grau BC, Horrocks I, Motik B, Parsia B, Patel-Schneider P, Sattler U. OWL 2: the next step for OWL. Web Semant Sci Serv Agents World Wide Web. 2008;6:309–22.

[87] Malone J, Holloway E, Adamusiak T, Kapushesky M, Zheng J, Kolesnikov N, et al. Modeling sample variables with an Experimental Factor Ontology. Bioinformatics. 2010;26:1112–8.

[88] Fernández JD, Martínez-Prieto MA, Gutierrez C. Compact Representation of Large RDF Data Sets for Publishing and Exchange. Semantic Web – ISWC 2010 [Internet]. Springer, Berlin, Heidelberg; 2010 [cited 2018 Mar 6]. p. 193–208. Available from: https://link.springer.com/chapter/10.1007/978-3-642-17746-0_13.

[89] Gray AJ, Goble C, Jimenez RC. Bioschemas: from potato salad to protein annotation. 2017.

[90] Breeding API [Internet]. [cited 2018 Mar 6]. Available from: https://brapi.org/.