Database, 2019, 1–11
doi: 10.1093/database/baz106
Original articles

Original articles

Enabling semantic queries across federated bioinformatics databases

Ana Claudia Sima[1,2,3,4,†], Tarcisio Mendes de Farias[2,3,4,5,†,*], Erich Zbinden[1,4], Maria Anisimova[1,4], Manuel Gil[1,4], Heinz Stockinger[4], Kurt Stockinger[1], Marc Robinson-Rechavi[4,5,*] and Christophe Dessimoz[2,3,4,6,7,*]

[1]ZHAW Zurich University of Applied Sciences, Obere Kirchgasse 2, 8400 Winterthur Switzerland, [2]Department of Computational Biology, University of Lausanne, 1015 Lausanne, Switzerland, [3]Center for Integrative Genomics, University of Lausanne, 1015 Lausanne, Switzerland, [4]SIB Swiss Institute of Bioinformatics, 1015 Lausanne, Switzerland, [5]Department of Ecology and Evolution, University of Lausanne, 1015 Lausanne, Switzerland, [6]Department of Genetics, Evolution, and Environment, University College London, Gower St, London WC1E 6BT, UK and [7]Department of Computer Science, University College London, Gower St, London WC1E 6BT, UK

*Corresponding author: Tarcisio.Mendesdefarias@unil.ch, Marc.Robinson-Rechavi@unil.ch, Christophe.Dessimoz@unil.ch
†These authors contributed equally to this work.

Citation details: Sima,A.C.,Mendes de Farias,T., Zbinden,E. *et al.* Enabling semantic queries across federated bioinformatics databases. *Database* (2019) Vol. 2019: article ID baz106; doi:10.1093/database/baz106

Received 28 June 2019; Revised 1 August 2019; Accepted 2 August 2019

Abstract

Motivation: Data integration promises to be one of the main catalysts in enabling new insights to be drawn from the wealth of biological data available publicly. However, the heterogeneity of the different data sources, both at the syntactic and the semantic level, still poses significant challenges for achieving interoperability among biological databases.

Results: We introduce an ontology-based federated approach for data integration. We applied this approach to three heterogeneous data stores that span different areas of biological knowledge: (i) Bgee, a gene expression relational database; (ii) Orthologous Matrix (OMA), a Hierarchical Data Format 5 orthology DS; and (iii) UniProtKB, a Resource Description Framework (RDF) store containing protein sequence and functional information. To enable federated queries across these sources, we first defined a new semantic model for gene expression called GenEx. We then show how the relational data in Bgee can be expressed as a virtual RDF graph, instantiating GenEx, through dedicated relational-to-RDF mappings. By applying these mappings, Bgee data are now accessible through a public SPARQL endpoint. Similarly, the materialized RDF data of OMA, expressed in terms of the Orthology ontology, is made available in a public

SPARQL endpoint. We identified and formally described intersection points (i.e. virtual links) among the three data sources. These allow performing joint queries across the data stores. Finally, we lay the groundwork to enable nontechnical users to benefit from the integrated data, by providing a natural language template-based search interface.

Database URL: http://biosoda.expasy.org and https://github.com/biosoda/bioquery

Introduction

One key promise of the postgenomic era is to gain new biological insights by integrating different types of data (e.g. 1, 2). For instance, by comparing disease phenotypes in humans with phenotypes produced by particular mutations in model species, it is possible to infer which human genes are involved in the disease (3).

A wealth of biological data is available in public data repositories; >100 key resources are featured in the yearly *Nucleic Acids Research* annual database issue (4). However, these databases vary in the way they model their data (e.g. relational, object-oriented or graph database models), in the syntaxes used to represent or query the data (e.g. markup or structured query languages) and in their semantics. This heterogeneity poses challenges to integrating data across different databases.

Ontologies have been widely used to achieve data integration and semantic data exchange (5–12). In this paper, by ontology, we adopt the broadly accepted definition in data and knowledge engineering of 'a formal, explicit specification of a shared conceptualization' (13). The relevance of ontologies in life sciences can be illustrated by the fact that repositories such as BioPortal (14) contain >700 biomedical ontologies, and the OBO Foundry (15) >170 ontologies. Moreover, major life sciences databases use ontologies to annotate and schematize data, such as UniProt (16) or ChEBI (17). Ontologies are important to enable knowledge sharing.

Currently, however, even when resources describe their data with ontologies, aligning these ontologies and combining information from different databases remain largely manual tasks, which require intimate knowledge of the way the data is organized in each source. This is despite a plethora of existing literature on data integration approaches, in particular in biological research (surveys of these approaches, as well as the challenges involved, include (18–20)). Projects such as KaBOB (21), Bio2RDF (22) and Linked Life Data (23) link different life science resources using a common ontology and data conventions. However, their centralized architecture makes it difficult to remain up-to-date and to scale up. For example, when querying the number of UniProt protein entries over the outdated and centralized Linked Life Data approach, we can only count ∼10% of the 230 million entries that are in the current UniProt release (see Supplementary data Section S1 for further explanations). To avoid this issue, federated approaches have recently been proposed (24–27), but to the best of our knowledge, none of them proposes a vocabulary and patterns to extensively, explicitly and formally describe how the data sources can be interlinked further than mostly considering 'same as'-like mappings; in effect, they put the burden on the users to find out precisely how to write a conjunctive federated query. An emerging research direction entails automatically discovering links between datasets using Word Embeddings (28). We did not pursue this approach, given that it is computationally expensive and that for our study writing the relational-to-Resource Description Framework (RDF) mappings proved more straightforward. However, Word Embeddings would be important for the case of integrating more data sources for which the connecting links (join points) are not clearly known. Most of the existing federated approaches address the problem of multiple database models by explicitly converting and storing the data into the same type of storage engine in order to achieve data interoperability. This often implies data duplication, which complicates maintenance. Among the aforementioned federated approaches, we can highlight the approach in (26), which requires less human interventions to generate federated queries. Nonetheless, this approach was mostly designed for chemical substance data based on predefined 'same as' mappings and handcrafted query patterns. Moreover, when considering the generated SPARQL query examples, they are mostly disjunctive queries (i.e. union) rather than complex conjunctive queries (i.e. intersection), which are our main focus. As opposed to other federated systems, such as BioFed (24), we do not focus on benchmarking or improving the performance of the underlying federation engine. However, our experiments with federated queries on the integrated data corroborate existing studies in showing that federation engines exhibit significant performance degradation when processing queries that involve large intermediate result sets (29).

To address the problem of semantic, syntactic and data model heterogeneity, we propose an ontology-driven linked data integration architecture. We apply this architecture to build a system that federates three bioinformatics databases containing: evolutionary relationships among genes across species (OMA), curated gene expression data (Bgee) and biological knowledge on proteins (UniProt). In Supplementary data, we summarize the key data provided by Bgee, OMA and UniProt (Supplementary data Table S3). Each of the three databases uses a different technical approach to store information: a Hierarchical Data Format 5 (HDF5; http://www.hdfgroup.org/HDF5/) data store (DS) for OMA (30), a relational database for Bgee (31) and an RDF store for UniProt (16). Our main contribution is to enable researchers to jointly query (i.e. conjunctive queries) the three heterogeneous databases using a common query language, by introducing and leveraging 'virtual links' between the three sources. Furthermore, we show how relational data can be made interoperable with RDF data 'without' requiring the original relational data to be duplicated into an RDF storage engine. This can be achieved by constructing dedicated relational-to-RDF mappings, allowing the unmodified original data to be queried via the structured query language SPARQL (32). In our proposed architecture, we illustrate this through the example of the Bgee relational database.

Moreover, for the purpose of building the federated data access system, we make the following additional contributions: (i) a semantic model for gene expression, (ii) an extension and adaptation of the Vocabulary of Interlinked Datasets (VoID) (33), (iii) public SPARQL 1.1 (32) query endpoints for OMA and Bgee and (iv) a user-friendly search interface based on an extensible catalogue of query templates in plain English. The main purpose of (iv) is to demonstrate that the different database models can be jointly queried based on our approach, but our system supports any general-purpose query builder compliant with SPARQL version 1.1, such as in (34–39).

Our article is structured as follows. In Section 2, we describe the individual databases, as well as a high-level introduction to our approach and the semantic models used in this work. In Section 3, we provide the implementation details of the three layers of our proposed architecture (DS, structured query interface (SQI) and application). In Section 4, we evaluate the performance of the system on a catalogue of 12 representative federated biological queries. Finally, we conclude with a discussion and outlook.

Materials and Methods

System Design

To understand more concretely the problem of integrating data from multiple sources, consider the following motivating example: 'What are the human genes which have a known association to glioblastoma (a type of brain cancer) and which furthermore have an orthologous gene expressed in the rat's brain?'. To answer this question, we would need to integrate information currently found in different databases:

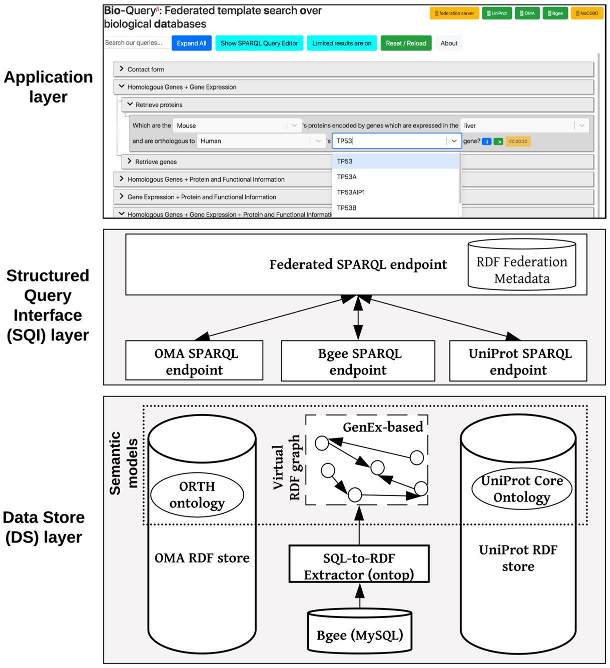
1. Human proteins associated with glioblastoma can be obtained from the UniProt KnowledgeBase, a database providing a comprehensive, high-quality sequence and functional information on proteins (16). In the rest of the paper, we will use the name UniProt for readability.
2. The orthologs of these proteins in the rat can be obtained from OMA, a database of orthology inferences (30). Orthologs are genes in different species that evolved from a common ancestral gene by speciation. The orthologs are normally thought to retain the same function in the course of evolution. Other homology information, such as one-to-one orthology or paralogy, can be derived from the hierarchical orthologous groups (HOGs) data structure (11, 40).
3. The genes expressed in the rat brain can be obtained from Bgee, a database of curated gene expression patterns in animals (31). Bgee version 14.0 includes gene expression data for 29 species such as human, mouse or hedgehog. Currently, Bgee data are stored in a MySQL relational database (41).

In the following, we first provide a high-level description of our approach, then introduce the semantic models involved.

A federated, ontology-driven data integration approach

In order to achieve semantic interoperability between Bgee, OMA and UniProt, we have chosen a federated approach based on ontologies (Figure 1). The advantage of a federated approach is to avoid imposing a common global schema or meta-model on all data sources and to facilitate the integration of further resources in the future. In doing so, we avoid, for example, the fastidious and time-consuming task of maintaining a centralized, integrated knowledge base. Instead, we provide a homogeneous data access layer to query the heterogeneous data sources. This homogeneous layer is part of a new generation of federated databases, such as polystores (42), that provide seamless access to distinct data models of storage engines (e.g. MySQL and RDF stores). Unlike that approach, we do not seek to optimize query performance by transferring data on-the-fly between disparate storage engines (42) but rather focus on solving syntactic and semantic heterogeneities among data stores.

To solve the syntactic heterogeneity, we rely on a structured query language—SPARQL—as the homogenous syntax to query all the data (32). We favoured SPARQL 1.1 over alternatives because it is World Wide Web Consortium (W3C) compliant, and the data to be integrated are on the



Which are the Mouse 's proteins encoded by genes which are expressed in the liver
and are orthologous to Human 's TP53 gene?

Figure 1. Overview of the ontology-driven federated data integration architecture applied to Bgee, OMA and UniProt. The application layer depicts a web search interface with editable templates to jointly query the data stores. Available online at http://biosoda.expasy.org.

web; because it supports federated queries; and because one of our target data stores, UniProt, is already accessible through a SPARQL 1.1 endpoint, alongside a growing number of other biological databases (29). Indeed, although our initial prototype integrates data from Bgee, OMA and UniProt, we plan to extend the system to include more data sources in the future.

To reduce semantic heterogeneity among the databases, we rely on ontologies further described in Subsection 2.2, which are defined with the Web Ontology Language 2 (OWL 2) and thus based on the RDF model and syntax (https://www.w3.org/TR/owl2-overview/). RDF-based modelling decisions were taken to mitigate this heterogeneity when adopting ontological terms and instances to structure and represent the non-RDF data—OMA HDF5 and Bgee relational data. For example, considering the OMA, Bgee and UniProt databases, UniProt completely covers the other two with regard to the taxonomic lineage information for an organism. Therefore, we rely on UniProt classes and instance internationalized resource identifiers (IRIs) when representing and modelling taxonomy-related data in the OMA and Bgee RDF serializations. As a further example, we can also mention the representation of genes among these three data sources. OMA genes completely cover Bgee genes, but not all OMA genes have corresponding ones in UniProt and *vice versa*. Because of this, we decided to model Bgee genes the same way as in OMA, thereby facilitating interoperability between gene expression and orthology data. The next sections describe in more details the semantic models and the federated architecture proposed.

Semantic models

Ontology-based data integration requires as a preliminary step that each of the individual resources composing the federated system provide an explicit ontological description of their data. To minimize the need for semantic reconciliation—i.e. the process of identifying and resolving semantic conflicts (43), for example, by matching concepts from heterogeneous data sources (44)—we sought to rely as much as possible on existing ontologies when defining new semantic models.

Prior to our current work, among the three databases considered in this article, only UniProt provided an RDF representation of its data, as well as a SPARQL endpoint. The current UniProt RDF release comprises >55 billion triples and is based on the OWL 2 Full UniProt core ontology described in (45).

For the orthology data in OMA, we adopted the orthology (ORTH) ontology (46), which was recently devised by the Quest for Orthologs Consortium (47) as a common data schema for integrating orthology databases, such as OMA. We use ORTH to structure the OMA data, which is primarily stored in an HDF5 DS. Furthermore, during the conception of a second version of ORTH, design decisions such as the adoption of taxon-related terms from the UniProt ontology were made in order to enhance interoperability, enabling us to establish links among the data stores (Subsection 3.2.2). Therefore, the work presented in this article also contributed towards a new, improved version of the ORTH ontology, which is described in (1).

In the case of Bgee, representing the original data in RDF proved to be a challenge, due to a general lack of a comprehensive ontology to serve as a data schema for describing knowledge in the field of gene expression. This may seem surprising considering the ubiquity of gene expression analyses in molecular biology and the existence of multiple well-established resources for gene expression—not only Bgee, but also Expression Atlas (EA) (48), Genevestigator (49), or the Tissue Expression database (50). We note here that different gene expression databases often use distinct criteria to assert 'expressed in' or 'absent in' relations.

To the best of our knowledge, two semantic models currently exist as initial attempts to structure gene expression related data: the Relation Ontology (RO) (51) and the EA model (52). The RO defines only a few terms within the domain of gene expression and is not specifically designed for this knowledge domain. Notably, it contains 'expressed in' and 'expresses' relations. The EA defines a semantic model related to gene expression that mainly focuses on modelling the EA data itself and not the domain of gene expression generally. In this EA model, additional data interpretations (i.e. semantics) are not explicitly represented, such as a given gene 'is expressed' or 'lowly expressed' in some sample relative to others. Although it would be possible to obtain this information through a more complex query on the EA SPARQL endpoint, we lack an explicit representation, which would allow us to compare gene expression data from these different databases.

To provide a first step towards a general-purpose gene expression ontology, we drafted a new semantic model called GenEx. GenEx is aligned with the RO and EA models to facilitate interoperability with existing RDF stores. We also included semantic rules and terms to address (i) the representation of additional information related to gene expression, such as developmental stages, as well as 'absent in' and 'highly expressed' relations; and (ii) the trade-off between virtualization and materialization for the sake of query execution time and data storage. Furthermore, we reuse parts of the data schemas of the ORTH and UniProt core ontologies to provide (iii) the capacity to interoperate with other biological databases from different knowledge domains that are still relevant to the gene expression domain. For example, integrating orthology and gene expression data is relevant since we might want to predict gene expression conservation for orthologous genes. The draft GenEx is available online and documented in https://biosoda.github.io/genex/.

We stress that GenEx is currently in draft state. To become a standard, it needs to be endorsed and supported by multiple key stakeholders. We plan to initiate discussions with representatives of Bgee, EA, Genevestigator and Tissue database teams and intend to solicit involvement from others, for example, the Model Organisms Databases (http://www.alliancegenome.org).

Implementation

Our federated data integration architecture comprises three layers: the DS layer, the SQI layer and the application layer (Figure 1). The DS layer contains all data stores to be integrated, including ontologies and methods to solve semantic and data model heterogeneities, such as relational-to-RDF mappings (Section 3.1). The SQI layer provides a homogeneous query language syntax and exploits common instances and literals (i.e. virtual links) to retrieve data from the DS layer (Section 3.2). The application layer includes any software tool that accesses the data stores through the SQI layer, for example, a web search interface (Section 3.3). Figure 1 illustrates this architecture applied to our use case: the Bgee, OMA and UniProt databases.

The three layers are described in the next subsections, and source code is available at https://github.com/biosoda/bioquery.

Data Store layer

The UniProt data were already available in an RDF model and accessible through a SPARQL endpoint at the start of our project. Therefore, we could use UniProt data as is.

The core of our work on the DS layer consisted in exposing data from Bgee and OMA as RDF, with the goal of solving data model heterogeneity. We focused our efforts on including the domain-specific, most 'value-added' aspects of Bgee and OMA to the DS layer—leaving out information already available in UniProt. As a result, the Bgee and OMA data accessible through our system are subsets of their original contents. We provide an overview of the types of information available in the original sources versus in their RDF representation in the Supplementary data (Table S3). This reduced the development work and data duplication among the databases, without loss of information considering that our federated approach enables directly retrieving this data from its original source (i.e. UniProt).

The Bgee data are stored in a relational database, meaning that integration between RDF stores and relational databases would still require substantial effort. There are two main methods to overcome this issue. First, the existing data could be represented entirely as RDF, which consequently would replace the relational model. A second approach would be to express the existing relational data as a virtual RDF graph, defined over ontological concepts and relations. We have chosen the latter approach, also referred to as 'ontology based data access' (OBDA) (53). Our choice is justified by the fact that changing the Bgee DS into an RDF model would either lead to data duplication or would require significant changes in the current Bgee analysis pipeline (31). This is because Bgee is now adapted to the relational model for storing raw and preprocessed data from multiple data sources such as Ensembl, GEO, ArrayExpress and others (https://bgee.org/?page=source).

To implement OBDA over the Bgee relational database, we used the Ontop platform (53) version 3.0-beta-2. We defined several relational-to-RDF model mappings, which dynamically instantiate the gene expression semantic model described in Subsection 3.1. Figure 2 shows a simplified example of OBDA mappings that serve to express data from the relational model in the RDF model. Namespace prefixes such as 'up:' shown in Figure 2 and used in the rest of this article are defined in Supplementary data Table S1. Some of the mappings can be simple 1-to-1 correspondences—for example, a gene name (shown in red colour on the right) can directly be used as a label of a 'orth:Gene' class instance. Other mappings require transforming the original attributes in the relational data for interoperability—for example, replacing ':' with '_' in the case of anatomical entity identifiers from Bgee to be compliant with the existing UBERON ontology IRI terms (54), as shown in green colour with

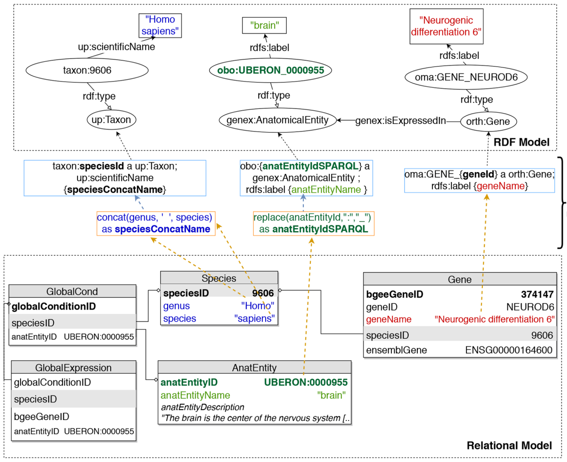

"Homo sapiens"
up:scientificName
taxon:9606
rdf:type
up:Taxon
"brain"
rdfs:label
obo:UBERON_0000955
rdf:type
genex:AnatomicalEntity
genex:isExpressedIn
"Neurogenic differentiation 6"
rdfs:label
oma:GENE_NEUROD6
rdf:type
orth:Gene
RDF Model
taxon:speciesId a up:Taxon; up:scientificName {speciesConcatName}
obo:{anatEntityIdSPARQL} a genex:AnatomicalEntity ; rdfs:label {anatEntityName }
oma:GENE_{geneId} a orth:Gene; rdfs:label {geneName}
OBDA mappings
concat(genus, ' ', species) as speciesConcatName
replace(anatEntityId,":","_") as anatEntityIdSPARQL
GlobalCond
globalConditionID
speciesID
anatEntityID UBERON:0000955
Species
speciesID 9606
genus "Homo"
species "sapiens"
Gene
bgeeGeneID 374147
geneID NEUROD6
geneName "Neurogenic differentiation 6"
speciesID 9606
ensemblGene ENSG00000164600
GlobalExpression
globalConditionID
speciesID
bgeeGeneID
anatEntityID UBERON:0000955
AnatEntity
anatEntityID UBERON:0000955
anatEntityName "brain"
anatEntityDescription
"The brain is the center of the nervous system [..
Relational Model

Figure 2. An illustration of relational-to-RDF mappings on a sample of the Bgee database. These mappings address both 'schema-level' heterogeneity (an example is shown in blue), as well as 'data-level' heterogeneity (shown in green). A mapping can also be a simple 1-to-1 correspondence between a relational attribute (e.g. 'geneName', shown in red) and its equivalent RDF property (in this case, an 'rdfs:label' of an 'orth:Gene' instance). Namespace prefixes are defined in Supplementary data Table S1.

the example of 'UBERON:0000955' in Figure 2. Another type of transformation can be even combining multiple columns to instantiate a concept, as in the case of expressing 'species' data from Bgee in terms of instances of 'up:Taxon'. In this case, the OBDA mapping serves to concatenate the 'genus' and 'species' columns from Bgee in order to form the scientific name in compliance with the UniProt taxonomy. The scientific names of species in UniProt are denominated through the 'up:scientificName' property, composed of both genus and species. This is illustrated in the left-most set of mappings (in blue colour) in Figure 2. For further details of this OBDA mapping, see the Supplementary data (Section S4). The full set of OBDA mappings used to expose Bgee relational data as virtual RDF triples are provided in https://github.com/biosoda/bioquery.

The code fragment in Listing 1.1 illustrates a mapping expressed with the Ontop relational-to-RDF mapping syntax, where the 'source' is a SQL SELECT statement and the 'target' consists of the corresponding RDF-based properties and classes. While direct and simple mappings (∼80% of the total) could in principle be automatically generated, complex ones such as the 'isExpressedIn' relationship shown in Listing 1.1 can only be manually defined. Further explanations about this are available in Supplementary data (Section S4).

Once relational-to-RDF mappings have been defined with Ontop between the Bgee MySQL database and GenEx, the original data can be queried with SPARQL, through the Bgee RDF virtual model. At query time, Ontop will translate SPARQL queries into SQL on-the-fly, using the mappings,

```
target  oma:GENE_{geneId} genex:isExpressedIn uberon:{
      anatEntityIdSPARQL} .
source  SELECT g.geneId,
        REPLACE(gc.anatEntityId,":","_") AS anatEntityIdSPARQL
        FROM globalExpression AS ge
        JOIN globalCond AS gc
        ON ge.globalConditionId = gc.globalConditionId
        JOIN gene AS g ON g.bgeeGeneId = ge.bgeeGeneId
```

Listing 1.1. Ontop mapping to infer the 'is expressed in' GenEx relation (i.e. target schema) based on the Bgee relational database (i.e. data source). Prefixes are defined in Supplementary data Table S1.

and execute these over the Bgee relational database. Ontop has the advantage of supporting federated queries as part of SPARQL 1.1 and of being open source. In order to enable researchers to directly use the RDF representation of Bgee, we made available a public SPARQL 1.1 endpoint at http://biosoda.expasy.org/rdf4j-server/repositories/bgeelight as a query service (without any webpage associated with it). Nonetheless, the OBDA solution with Ontop has some limitations—we discuss some of these in Section S4 in the Supplementary data.

The OMA data are internally stored in an HDF5 file. This is not a database management system such as MySQL but rather a data model and file format along with an application programming interface (API), libraries and tools. Similarly to Bgee, we have to homogenize the OMA database model and syntax in order to enable integration with other biological RDF DSs (either virtual or materialized). For OMA, we chose to materialize the key parts of OMA data as an RDF graph, by implementing a hybrid approach, that combines materialization and a possible RDF graph virtualization for the sake of semantic enrichment and knowledge extraction, as described in detail in https://qfo.github.io/OrthologyOntology. The OMA RDF data and ORTH ontology are stored in a Virtuoso 7.2 triple store, and a SPARQL endpoint is available at https://sparql.omabrowser.org/sparql. Further explanations regarding the OMA RDF data materialization are available in Supplementary data Section S5.

SQI layer

Once the data stores are accessible through SPARQL endpoints, as depicted in Subsection 3.1, we can exploit means to link them at the data level. To do so, we identify common class instances and literals (e.g. strings) in order to establish 'virtual links'. We define a virtual link as an intersection data point between two data stores. The links are required in order to enable performing federated queries, given that they act as join points between the federated sources. Figure 3 illustrates virtual links among UniProt, Bgee and OMA. For example, OMA and Bgee describe complementary

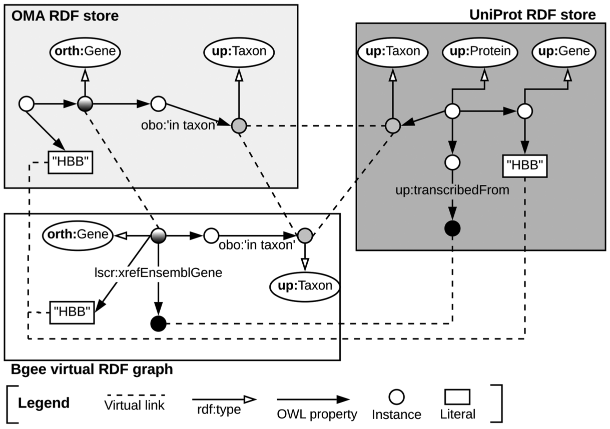


Figure 3. Example of virtual links among UniProt, OMA and Bgee data stores.

information about common genes (instances of the 'orth:Gene' class), as well as taxa (instances of the 'up:Taxon' class), both of which can serve as virtual links to connect the two sources. A federated SPARQL query written based on the virtual links is described in Supplementary data Section S3. To formally and explicitly describe virtual links, we adapted and extended the VoID RDF schema vocabulary (33) to include the concept of virtual links. We call this vocabulary Extended VoID (VoIDext). VoIDext is fully specified and exemplified in https://biosoda.github.io/voidext/. The entire metadata of virtual links among UniProt, Bgee and OMA RDF stores for the work depicted in this article are available at http://purl.org/query/bioquery. In the VoIDext specification, we also depict the SPARQL queries to retrieve the virtual links among OMA, Bgee and UniProt that support the writing of joint federated SPARQL queries. These queries can be executed on the federated SPARQL endpoint illustrated in Figure 1 http://biosoda.expasy.org:8890/sparql.

To leverage virtual links between Bgee and the other databases, we took advantage of the flexibility provided by Ontop when defining the Bgee OBDA mappings. We aimed at mapping Bgee data into corresponding instance IRIs and literals that already exist in the OMA and UniProt RDF graphs. For example, species-related instance IRIs in the Bgee virtual graph are indeed exact matches of 'up:Taxon' instance IRIs that are stored in the UniProt database.

Likewise, the code fragment in Listing 1.1 asserts the (re)use of OMA gene instances as part of the Bgee virtual RDF graph rather than creating new Bgee ones. In this way, we avoid additional 'owl:sameAs' assertions to state that the two instances are actually the same. Thus, 'orth:Gene' instances are intersection nodes (i.e. virtual links) between the Bgee and OMA graphs. Figure 3 (left-hand side) illustrates a shared 'orth:Gene' instance between OMA and Bgee graphs. Further information about the virtual links depicted in Figure 3 is available in Section S6 in the Supplementary data.

Overall, we provide a federated SPARQL query endpoint along with an RDF store that exclusively contains metadata about the virtual links and the SPARQL endpoints of UniProt, OMA and Bgee data stores. These metadata based on the VoIDext schema precisely define and document how the distributed datasets can be interlinked. Therefore, they may significantly facilitate the manual or automatic writing of a SPARQL 1.1 federated query, given that users are no longer required to discover the interlinks between the queried datasets on their own. In (55), we detail the drawbacks of the current VoID link sets to represent virtual links and the description of the novel VoIDext specification.

Application layer

The main goal of the application layer in our work is to enable users, even with no prior technical training, easy access to the integrated information from the three biological databases. We developed a user-friendly interface (illustrated in the top part of Figure 1), which is accessible at http://biosoda.expasy.org/. The interface presents a catalogue of representative query templates drafted together with domain experts. The queries are provided in natural language, with editable fields, and grouped in a tree structure according to the target knowledge domain(s) and information retrieved for each query. A search bar is also provided, which enables filtering the templates by keywords of interest (e.g. 'disease').

For users with more advanced technical expertise, we also provide the option to show and modify the equivalent SPARQL queries. In doing so, our approach has the potential to increase the productivity of domain scientists in exploring the three heterogeneous datasets jointly. Additionally, the catalogue of questions is destined to grow according to user needs and feedback.

Moreover, because of the federated architecture of our system, its performance depends on that of the underlying data sources, e.g. UniProt. The availability of the underlying data stores is indicated by green labels in the top right corner of the web page. For unavailable sources, the corresponding label is shown in yellow, as illustrated in the application layer in Figure 1.

By default, our system limits the total number of results returned, which allows for a faster response—the estimated response time is shown as a tag next to each query. However, the user can turn the limit option off, in order to obtain the full set of results. In this case, the response time may be significantly higher, which can largely be attributed to the SPARQL query execution time on the underlying sources. This has already been noted in similar previous systems

(56). In terms of scalability, the UniProt SPARQL endpoint is a good example, having already an active user base of >1000 users per month. As we directly rely on the infrastructure of the underlying sources, we can therefore expect our system to exhibit reasonable performance for approximately the same number of users.

Results

In this section, we first revisit our motivating example for integrated data access to the three databases (UniProt, OMA and Bgee) and then present experimental results based on a catalogue of 12 federated queries. All results are reproducible through our public interface described in Subsection 3.3.

Recall our motivating example from the start of Section 2: 'What are the human genes which have a known association to glioblastoma (a type of brain cancer) and which furthermore have an orthologous gene expressed in the rat's brain'. Answering the question requires solving the following three subqueries:

1. Retrieve human proteins with a disease description related to glioblastoma from UniProt.
2. Retrieve orthologs of these proteins in the rat from OMA.
3. Only keep those orthologs for which there exists evidence of expression in the rat's brain from Bgee.

The above steps translate to the federated SPARQL query in Supplementary data Listing S2. The query produces a set of 15 human–rat orthologous pairs and can be executed in any SPARQL 1.1 endpoint. The full query, as well as a detailed list of results, is available in the Supplementary data in Section S3.

We further evaluated the performance, in terms of runtime, of 12 federated queries that illustrate real use cases requiring information across the three databases (Table 1). The results are reproducible through our public template-based search interface. A detailed analysis of the queries, including their natural language description, the equivalent federated SPARQL queries, as well as an explanation of the complexity for each query, can be found at https://github.com/biosoda/bioquery.

Table 2 shows that most of the queries can be executed in a few seconds—up to 6 s for 9 out of 12 queries, with less than half a second for 3 out of these. This holds even for queries with higher complexity (number of triple patterns). A triple pattern is similar to a regular RDF triple, except that any part of the triple can be replaced by a variable (32). Although preliminary, the results in Table 2 are encouraging for the use of SPARQL queries in data exploration tasks or in an interactive environment.

The outlier Q10 calls for discussion. By comparing the natural language description of Q10 against Q11 (see corresponding entries in Table 1, where the difference between the two queries is highlighted in bold in the description of Q11), we can intuitively deduce that the complexity stems from the high degree of generality of the subquery that targets OMA. In the case of Q10, retrieving an answer will require scanning the entire available orthology data and retrieving a large intermediate result set (orthologs found in any species, a total of 2269 results). By contrast, Q11 restricts the search space to the 'primates' taxon only, which in practice results in a much lower query execution time (and a total of only 81 results). An important lesson derived from this example is that queries should always be as specific as possible, in order to limit both the search space and the size of intermediate results to the minimum necessary to obtain a relevant answer. Although this query illustrates a worst-case scenario, the results are still returned in <6 min—a latency that is tolerable for investigations in a biological research context.

Discussion and outlook

Data integration across heterogeneous biological databases promises to be one of the catalysts for gaining new biological insights in the postgenomic era. Here, we introduced an ontology-driven approach to bioinformatic resource integration. This approach enables complex federated queries across multiple domains of biological knowledge, such as gene expression and orthology, without requiring data duplication. The integration of the three sources promises to open the path for novel comparative studies across species, for example, through the analysis of orthologs (OMA) of human disease-causing genes (UniProt) and their expression patterns in model organisms (Bgee). Thanks to modelling decisions made at the semantic (ontology) and data (assertions) levels, we established various virtual links among Bgee, OMA and UniProt. Moreover, making these virtual links available in VoIDext facilitates the task of writing federated SPARQL queries, since users have an explicit representation of the connections (join points) between the three data sources. We furthermore lay the groundwork for bringing the benefits of integrated data to domain specialists through a template-based search engine available online, which does not require users to know SPARQL in order to pose questions on the integrated data.

The catalogue of federated queries across the three data sources can serve as a starting point towards answering new biological questions that span across the domains of evolutionary relationships and gene expression. The results presented in this study can be easily reproduced through our template search interface. We furthermore make available

Table 1. Descriptions of the 12 federated queries across OMA, Bgee, UniProt used for evaluating our system; the queries can be further refined and executed through our template search interface available at http://biosoda.expasy.org/

Query	Description
Q1	Proteins in OMA encoded by the INS gene and their evidence type from UniProt.
Q2	Rabbit proteins encoded by genes orthologous to the HBB-Y gene in mouse and their associated information from UniProt.
Q3	*Rattus norvegicus* proteins paralogous to Tp53 and their UniProt function annotations.
Q4	Mouse genes expressed in the liver that are orthologous to the human INS gene.
Q5	Genes orthologous to a gene expressed in the fruit fly brain.
Q6	Genes in primates orthologous to a gene expressed in the fruit fly brain.
Q7	Anatomic entities where the ins zebrafish gene is expressed and the ins gene GO annotations.
Q8	Genes expressed in the human pancreas and their annotation in UniProt.
Q9	Genes expressed in the human brain during the infant stage and their UniProt disease annotation.
Q10	The orthologs of a gene that is expressed in the fruit fly brain and the UniProt annotations of these orthologs.
Q11	The orthologs in primates of a gene that is expressed in the fruit fly brain, and the UniProt annotations of the primate orthologs.
Q12	Proteins in humans, with a disease annotation, that are orthologous to a gene expressed in the rat brain.

Table 2. Tests performed to evaluate our approach in terms of query execution time and the number of results. We evaluated 12 federated queries of varying complexity (measured in terms of number of triple patterns). Their description is provided in Table 1. All queries were executed twenty times, providing an average runtime and its standard deviation, given in seconds. The longest running query, Q10, is highlighted in bold.

Query	Sources	#Results	Mean run-time (s)	SD run-time (s)
Q1	OMA, UniProt	27	5.13	0.13
Q2	OMA, UniProt	3	0.36	0.02
Q3	OMA, UniProt	1	0.47	0.05
Q4	OMA, Bgee	2	0.37	0.05
Q5	OMA, Bgee	5322	4.72	0.2
Q6	OMA, Bgee	38	2.38	0.09
Q7	Bgee, UniProt	16	68.18	105.85
Q8	Bgee, UniProt	58	33.17	25.23
Q9	Bgee, UniProt	6	2.37	0.04
Q10	**Bgee, OMA, UniProt**	**2269**	**349.18**	**4.19**
Q11	Bgee, OMA, UniProt	81	6.4	0.14
Q12	Bgee, OMA, UniProt	3	5.24	0.11

all source code, including the template search interface code, relational-to-RDF mappings and the catalogue of queries, with the goal of facilitating reuse of these components for further research. All resources are available in our GitHub repository.

Our experiments show that most queries in our catalogue can be answered within seconds. And although the more complex queries take several minutes to complete, we expect this turnaround time to be tolerable for most interested users—particularly considering the alternative of manually querying the resources and combining the results. In the future, we plan to include more resources in the federated system, focusing primarily on publicly available databases. We plan to start with those that already provide SPARQL 1.1 endpoints, for which the main work would entail defining virtual links to our existing integrated resources. In a second step, we can envision integrating more relational databases, for which the main work required would be to define the relational-to-RDF mappings, analogous to those presented for Bgee in the current work. Our first aim is to make more of the publicly accessible databases interoperable for the purpose of advancing research through integrated data access. Nevertheless, we can envision also integrating access control policies in the future, which would enable including sensitive resources, such as patient databases, in the federated system. A good starting point for understanding the types of existing access controls for RDF data, in order to accommodate these in our federation architecture, is the recent survey (57). Finally, we plan to add a federated query optimizer to our system to further improve the response time. We also note here that the application interface directly queries the underlying databases without performing additional tasks, such as considering all gene name synonyms to get broader results. We plan to support such features as part of future work. To support virtual link evolution, we aim to develop a tool to automatically detect broken virtual links because of either data schema changes or radical modifications of instances' IRIs and property assertions. Meanwhile, we encourage contributions to the current query catalogue, which will serve in the study of a natural language search interface for the integrated biological data as part of future work.

Supplementary data

Supplementary data are available at *Database* Online.

Funding

Swiss National Research Programme 75 'Big Data' (167149); Swiss National Science Foundation (150654); Swiss State Secretariat for Education, Research and Innovation (SIB Infrastructure grants) for OMA, Bgee, and Core-IT.

Conflict of interest. None declared.

References

1. Ritchie,M.D., Holzinger,E.R., Li,R.*et al.* (2015) Methods of integrating data to uncover genotype-phenotype interactions. *Nat. Rev. Genet.*, **16**, 85–97.
2. Karczewski,K.J. and Snyder,M.P. (2018) Integrative omics for health and disease. *Nat. Rev. Genet.*, **19**, 299–310.
3. Mungall,C.J., McMurry,J.A., Köhler,S.*et al.* (2017) The monarch initiative: an integrative data and analytic platform connecting phenotypes to genotypes across species. *Nucleic Acids Res.*, **45**, D712–D722.
4. Rigden,D.J. and Fernández,X.M. (2018) The 2018 nucleic acids research database issue and the online molecular biology database collection. *Nucleic Acids Res.*, **46**, D1–D7.
5. Zhang,H., Guo,Y., Li,Q.*et al.* (2018) An ontology-guided semantic data integration framework to support integrative data analysis of cancer survival. *BMC Med. Inform. Decis. Mak.*, **18**, 41.
6. Baitaluk,M. and Ponomarenko,J. (2010) Semantic integration of data on transcriptional regulation. *Bioinformatics*, **26**, 1651–1661.
7. Wang,C., Ma,X. and Chen,J. (2018) Ontology-driven data integration and visualization for exploring regional geologic time and paleontological information. *Comput. Geosci.*, **115**, 12–19.
8. Farias,T.M., Roxin,A. and Nicolle,C. (2015) FOWLA, a federated architecture for ontologies. *Rule Technologies: Foundations, Tools, and Applications*. Springer, Cham International Publishing, pp. 97–111.
9. Mate,S., Köpcke,F., Toddenroth,D. (2015) Ontology-based data integration between clinical and research systems. *PLoS One*, **10**, e0116656.
10. Knoblock,C.A. and Szekely,P. (2015) Exploiting semantics for big data integration. *AI Magazine*, **36**: 25–38. doi: 10.1609/aimag.v36i1. 2565.
11. de Farias,T.M., Chiba,H. and Fernández-Breis,J.T. (2017) *Leveraging Logical Rules for Efficacious Representation of Large Orthology Datasets*.
12. Shoaib,M., Ansari,A.A. and Ahn,S.-M. (2017) Cmapper: gene-centric connectivity mapper for EBI-RDF platform. *Bioinformatics*, **33**, 266–271.
13. Studer,R., Benjamins,V.R. and Fensel,D. (1998) Knowledge engineering: principles and methods. *Data Knowl. Eng.*, **25**, 161–197.
14. Whetzel,P.L., Noy,N.F., Shah,N.H.*et al.* (2011) BioPortal: enhanced functionality via new web services from the national center for biomedical ontology to access and use ontologies in software applications. *Nucleic Acids Res.*, **39**, W541–W545.
15. Smith,B., Ashburner,M., Rosse,C. *et al.* (2007) The OBO foundry: coordinated evolution of ontologies to support biomedical data integration. *Nat. Biotechnol.*, **25**, 1251–1255.
16. UniProt Consortium. (2018) UniProt: the universal protein knowledgebase. *Nucleic Acids Res.*, **46**, 2699.
17. Hastings,J., de Matos,P., Dekker,A. *et al.* (2013) The ChEBI reference database and ontology for biologically relevant chemistry: enhancements for 2013. *Nucleic Acids Res.*, **41**, D456–D463.
18. Goble,C. and Stevens,R. (2008) State of the nation in data integration for bioinformatics. *J. Biomed. Inform.*, **41**, 687–693.
19. Zhang,Z., Bajic,V.B., Jun,Y. *et al.* (2011) Data integration in bioinformatics: current efforts and challenges. *Bioinformatics-Trends and Methodologies Mahmood A. Mahdavi*. IntechOpen: 41–56. DOI: 10.5772/21654.
20. Lapatas,V., Stefanidakis,M., Jimenez,R.C. *et al.* (2015) Data integration in biological research: an overview. *J. Biol. Res. (Thessalon.)*, **22**, 9.
21. Livingston,K.M., Bada,M., Baumgartner,W.A., Jr *et al.* KaBOB: ontology-based semantic integration of biomedical databases. *BMC Bioinformatics*, **16**, 126.
22. Belleau,F., Nolin,M.-A., Tourigny,N. *et al.* (2008) Bio2rdf: towards a mashup to build bioinformatics knowledge systems. *J. Biomed. Inform.*, **41**, 706–716.
23. Momtchev,V., Peychev,D., Primov,T. *et al.* (2009) Expanding the pathway and interaction knowledge in linked life data. *Proceedings of International Semantic Web Challenge ISWC 2009 Chantilly*, Springer Lecture Notes in Computer Science LNCS VA, USA, **5823**. doi: 10.1007/978-3-642-04930-9.
24. Hasnain,A., Mehmood,Q., e Zainab,S.S. *et al.* (2017) Biofed: federated query processing over life sciences linked open data. *J. Biomed. Semantics*, **8**, 13.
25. Djokic-Petrovic,M., Cvjetkovic,V., Yang,J. *et al.* (2017) Pibas fedsparql: a web-based platform for integration and exploration of bioinformatics datasets. *J. Biomed. Semantics*, **8**, 42.
26. Arsić,B., Dokić-Petrović,M, Spalević,P, *et al.* (2019) SpecINT: a framework for data integration over cheminformatics and bioinformatics RDF repositories. Semantic Web Journal, **10**: 795–813. doi: 10.3233/SW-180327.
27. Wimalaratne,S.M., Bolleman,J., Juty,N. *et al.* (2015) SPARQL-enabled identifier conversion with identifiers.org. *Bioinformatics*, **31**, 1875–1877.

28. Fernandez,R.C., Mansour,E., Qahtan,A.A.*et al.* (2018) Seeping semantics: linking datasets using word embeddings for data discovery. *IEEE 34th International Conference on Data Engineering (ICDE) 2018,* Paris, France. IEEE, pp. 989–1000.
29. Saleem,M., Hasnain,A. and Ngonga Ngomo,A.-C. (2018) LargeRDFBench: a billion triples benchmark for SPARQL endpoint federation. *Web Semant.*, **48**, 85–125.
30. Altenhoff,A.M., Glover,M., Train,C.-M. (2018) The OMA orthology database in 2018: retrieving evolutionary relationships among all domains of life through richer web and programmatic interfaces. *Nucleic Acids Res.*, **46**, D477–D485.
31. Bastian,F., Parmentier,G., Roux,J.*et al.* (2008) Bgee: integrating and comparing heterogeneous transcriptome data among species. *Data Integration in the Life Sciences*, Lecture Notes in Computer Science, pp. 124–131. Springer Berlin Heidelberg.
32. Harris,S., Seaborne,A. and Prud'hommeaux,E. (2013) SPARQL 1.1 query language. *W3C Recommendation*, **21**, 778.
33. Alexander,K., Cyganiak,R., Hausenblas,M. *et al.* (2011) Describing linked datasets with the VoID vocabulary.
34. Hu, W., Qiu, H., Huang, J. *et al.* (2017) BioSearch: a semantic search engine for Bio2RDF. *Database (Oxford)*, **2017**, 10.1093/database/bax059.
35. De Leon Battista,A., Villanueva-Rosales,N., Palenychka,M.*et al.* (2007) SMART: a web-based, ontology-driven, semantic web query answering application. *Semantic Web Challenge*, **295**, 129–36
36. Dietze,H. and Schroeder,M. (2009) GoWeb: a semantic search engine for the life science web. *BMC Bioinformatics*, **10**, S7.
37. Bielefeldt,A., Gonsior,J. and Krötzsch,M. (2018) Practical linked data access via SPARQL: the case of wikidata. *Proceeding WWW2018 Workshop on Linked Data on the Web (LDOW-18)*. CEUR Workshop Proceedings, CEUR-WS.org. Lyon, France. ceur-ws.org.
38. García-Godoy,M.J., Navas-Delgado,I. and Aldana-Montes,J. (2012) Bioqueries: a social community sharing experiences while querying biological linked data. *Proceedings of the 4th International Workshop on Semantic Web Applications and Tools for the Life Sciences, SWAT4LS'11*. ACM, New York, NY, USA, pp. 24–31.
39. H. Chiba and I. Uchiyama. (2017) SPANG: a SPARQL client supporting generation and reuse of queries for distributed RDF databases. *BMC Bioinformatics*, **18**, 93.
40. Altenhoff,A.M., Gil,M., Gonnet,G.H. *et al.* (2013) Inferring hierarchical orthologous groups from orthologous gene pairs. *PLoS One*, **8**, e53786.
41. Komljenovic,A., Roux,J., Wollbrett,J. *et al.* (2018) BgeeDB, an R package for retrieval of curated expression datasets and for gene list expression localization enrichment tests. *F1000Res.*, **5**. DOI: 10.12688/f1000research.9973.2.
42. Gadepally,V., Chen,P., Duggan,J.*et al.* (2016) The BigDAWG polystore system and architecture. *IEEE High Performance Extreme Computing Conference (HPEC)*, Waltham, MA USA, pp. 1–6.
43. Siegel,M.D. and Madnick,S.E. (1991) *A Metadata Approach to Resolving Semantic Conflicts*, Cambridge, Massachusetts.
44. Gal,A., Modica,G., Jamil,H. *et al.* (2005) Automatic ontology matching using application semantics. *AI magazine*, **26**, 21.
45. Redaschi,N., Uniprot Consortium *et al.* (2009) Uniprot in RDF: tackling data integration and distributed annotation with the semantic web. Nature Precedings, 3rd Biocuration Conference, 2019, DOI: 10.1038/npre.2009.3193.1.
46. Tomás Fernández-Breis,J., Chiba,H., Del Carmen Legaz-García,M. *et al.* (2016) The orthology ontology: development and applications. *J. Biomed. Semantics*, **7**, 34.
47. Forslund,K., Pereira,C., Capella-Gutierrez,S. *et al.* (2017) Gearing up to handle the mosaic nature of life in the quest for orthologs. *Bioinformatics*, **34.2**, 323–329.
48. Petryszak,R., Burdett,T., Fiorelli,B. (2014) Expression atlas update—a database of gene and transcript expression from microarray- and sequencing-based functional genomics experiments. *Nucleic Acids Res.*, **42**, D926–D932.
49. Hruz,T., Laule,O., Szabo,G.*et al.* (2008) Genevestigator v3: a reference expression database for the meta-analysis of transcriptomes. *Adv. Bioinformatics*, **2008**, 420747.
50. Palasca,O., Santos,A., Stolte,C.*et al.* (2018) TISSUES 2.0: an integrative web resource on mammalian tissue expression. *Database (Oxford)*, **2018**. DOI: 10.1093/database/bay003.
51. Smith,B., Ceusters,W., Klagges,B. *et al.* (2005) Relations in biomedical ontologies. *Genome Biol.*, **6**, R46. ISSN: 1474-760X. doi:10.1186/gb-2005-6-5-r46.
52. Jupp,S., Malone,J., Bolleman,J. (2014) The EBI RDF platform: linked open data for the life sciences. *Bioinformatics*, **30**, 1338–1339.
53. Calvanese,D., Cogrel,B., Komla-Ebri,S. *et al.* (2017) Ontop: Answering SPARQL queries over relational databases, *Semantic Web*, IOS Press, **8**, 3, 471–487.
54. Mungall,C.J., Torniai,C., Gkoutos,G.V. *et al.* (2012) Uberon, an integrative multi-species anatomy ontology. *Genome Biol.*, **13**, R5. ISSN: 1474-760X. doi:10.1186/gb-2012-13-1-r5.
55. de Farias,T.M., Stockinger,K. and Dessimoz,C. (2019) VoIDext: Vocabulary and patterns for enhancing interoperable datasets with virtual links. arXiv preprint arXiv:1906.01950v2.
56. Ferré,S. (2017) Sparklis: an expressive query builder for sparql endpoints with guidance in natural language. *Semantic Web*, **8**, 405–418.
57. Kirrane,S., Mileo,A. and Decker,S. (2017) Access control and the resource description framework: a survey. *Semantic Web*, **8**, 311–352.